MODELS AND METHODS FOR COMPUTER SIMULATIONS AS A RESOURCE IN PLANT BREEDING

BY

XIAOCHUN SUN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Crop Sciences
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Associate Professor Rita H. Mumm, Chair
Professor Roger D. Shanks
Associate Professor Ping Ma
Assistant Professor Patrick J. Brown

Abstract

A number of crucial decisions face the plant breeder in developing improved cultivars. Because of the impact and modern complexity of these decisions, computer simulation based on effective models can be an important resource for the breeder, particularly in providing guidance on choice of parents and decisions related to various aspects of the integrated breeding approach. Four areas related to use of computer simulation and modeling were explored as outlined in respective chapters:

1) To maximize utility, the simulation tool must be based on effective models of the genome and the process of genetic transmission through generations, the breeding process, and other 'processes' involved in genetic recombination, identification and production of new cultivars. Additionally, the statistical methodology employed has ramifications for predicting performance and breeding outcome. We highlighted the role of computer simulation in planning phases of crop genetic improvement, the basics of model building, statistical considerations, and key issues to be addressed. Examples of publicly available simulation software were described (features, functionalities, and assumptions) and new directions for improved/expanded approaches and tools are discussed.

2) Improvement of genome model, through accurate modeling investigation of crossover interference and additive and dominance effects were explored. Crossover interference in maize was modeled by two-pathway methods using doubled haploid data; various levels of crossover interference were found across chromosomes. To challenge the commonly invoked assumption of the infinitesimal model of genetic effects, published data from five quantitative trait loci (QTL) mapping studies were used to derive the distributions of QTL additive effects and dominance coefficients in the form of mixtures of normals. Four separate normal distributions

with zero mean and different variances were fitted for QTL additive effects of different classes of traits. Dominance coefficients fit a single normal distribution with positive mean, indicating prevalence of additive or partial dominant gene action across many traits.

3) To enhance the predictive ability in new line development, we developed and evaluated a novel method for use in genomic selection. Genomic selection procedures have proven useful in estimating breeding value and predicting phenotype with genome-wide molecular marker information. We proposed a new nonparametric method, pRKHS, which combined the features of supervised principal component analysis and reproducing kernel Hilbert spaces regression, with versions for traits with no/low epistasis, pRKHS-NE, to high epistasis, pRKHS-E. Compared to RR-BLUP, BayesA, BayesB, and RKHS, pRKHS delivers greater predictive ability, particularly when epistasis impacts expression in the trait of interest. Beyond prediction, the new method also facilitated inferences about the extent to which epistasis influences trait expression.

4) A case study involving transgenic conversion of a target hybrid for 15 events explored optimization of parameters in version testing to facilitate a high likelihood of recovery of at least one hybrid conversion with performance equivalent to the unconverted target hybrid. We determined that by creating 5 versions of each parental conversion, then selecting 3 each based on breeding value and yield testing a total of 9 hybrid versions of the conversion facilitated a 95% probability of success. These results had implications for the trait conversion process pertaining to single event conversion and event pyramiding.

Acknowledgements

I am very grateful to my advisor Dr. Rita H. Mumm for her sincere guidance and support during my training at the Maize Genetics Lab.

I would like to thank my supervisory committee members Drs. Glenn R. Johnson, Roger D. Shanks, Ping Ma, and Patrick J. Brown for their valuable comments and advice.

I would also like to thank all technicians, post doctorate fellows and graduate and undergraduate students working at the Maize Genetics Lab for their kind technical assistance and friendship.

Finally, I would like to thank my family for their support throughout my PhD study.

Table of Contents

List of Figures

List of Tables

Chapter 1 - Introduction

1.1 Literature review of current simulation programs and models

1.1.1 Introduction

Crop cultivars with improved performance have been successfully developed through plant breeding, which comprises activities such as crossing, inbreeding, progeny selection, and seed propagation. Before any such activities are actually initiated, the breeder has several key decisions to make, including 1) choice of germplasm and specific parents, and 2) various options pertaining to the breeding strategy such as breeding methods. The prediction of phenotypic response to selection is quantified by genetic gain (ΔG), which is defined as the rate of change in the population mean under selection (Falconer and Mackay 1996; Moose and Mumm 2008):

$$\Delta G = h^2 \times \sigma_p \times \frac{i}{L} = \frac{\sigma_a^2 \times i}{\sigma_p \times L}$$

where h^2 stands for narrow sense heritability, σ_p stands for phenotypic standard deviation, i stands for selection intensity, L stands for generation interval, and σ_a^2 stands for additive variance. Thus, genetic gain is a function of these variables, some of which are impacted by choice of parents (e.g. σ_p and σ_a^2) and others are a consequence of the chosen breeding strategy (e.g. the values of i and L).

Choice of parents is foundational to cultivar improvement. The maxim of crossing "good by good" to produce a set of progenies from which a superior line will be derived assumes highly positive mean performance for the traits of interest and diversity in the favorable alleles contributed by the parents (Melchinger et al. 1988). Breeders have used several ways to assess these among potential parents, including phenotypic expression of heterosis, best linear unbiased

prediction (BLUP), estimates of genetic relationship based on pedigree, molecular marker data, etc. (Bernardo 2002; Burkhamer et al. 1998; Dudley 2004; Dudley et al. 1992; Frisch et al. 2010; Panter and Allen 1995; Zhong and Jannink 2007). Dudley (1982, 1987) proposed a theory for choosing parents to improve a single cross based on classes of loci. In this scenario, three inbreds, P_1, P_2 and P_w, are involved: P_w is used to cross with parent P_1 (or P_2) to develop a new inbred P_{new}, whose cross with P_2 (or P_1) would outperform the original cross P_1 x P_2. Methods for choosing P_w have been progressively improved: 1) Dudley (1984) suggested that P_w should have the largest number of favorable alleles not present in P_1 or P_2; 2) Bernardo (1990) further proposed a statistic method called "net improvement" to account for the risk of losing favorable alleles already present in P_1 x P_2 when introducing favorable alleles from P_w; and 3) Metz (1994) proposed using the probability of net gain of favorable alleles to estimate the likelihood that P_w could improve performance of a given single cross. Furthermore, P_w should be chosen to maintain the same heterotic pattern as P_1 to maximize heterosis if P_2 is used as tester, and vice versa. More recently, ways to estimate breeding value of potential parents have been utilized to direct choice of parents in crop improvement, e.g. BLUP, genomic selection (GS) or genome wide selection (GWS). In addition, choice of parents may be based on predicted performance. Zhong and Jannink (2007) proposed choosing inbred parents based on the expected performance of the best progeny resulting from that cross, referred to as 'superior progeny value'.

Choice among options pertaining to aspects of the breeding strategy employed is also critical to successful crop improvement. The breeding strategy entails more than simply the breeding methods used; it encompasses an integrated approach which also details the specific breeding objectives aligned with the overall goals, the technologies deployed, the experimental designs and plans for statistical analysis of the data collected, structures of the population to be

created, screening/phenotyping methods and devices, and selection thresholds and regimes. The breeder must consider the size of the effects of favorable alleles contributed by each parent (Bernardo and Yu 2007), the number of loci likely to be involved with the trait(s) under selection (Bernardo 2002), and the likely gene action underlying these (Cockerham 1954; Crow and Kimura 2009). Furthermore, modes of male sterility to be utilized for hybrid cultivars, how to manage selection for multiple traits (e.g. tandem selection, index selection), use of dihaploidy to accelerate inbreeding, and use of genomic applications like molecular markers must be considered in deciding the integrated approach or other methods of indirect selection (Moose and Mumm 2008). For the latter, population structure may play a key role in estimating marker effects, particularly if depending on linkage disequilibrium (LD) to identify markers in genomic proximity to genes underlying traits of interest (Caldwell et al. 2006; Hamblin et al. 2005; Hyten et al. 2007; Remington et al. 2001). Approaches may differ depending on whether cultivars are inbred or hybrid (Bernardo 2002; Kharkwal and Roy 2004) and whether transgenic traits are involved (Mumm 2007). To increase quantity as well as quality of phenotypic data, much effort has been devoted to field design optimization and automated data collection (Eathington et al. 2007; Hallauer and Pandey 2006) aimed at greater accuracy and precision in assessing performances.

The decisions made by the breeder with respect to choice of parents and breeding strategy can mean the difference between success and failure in meeting the defined breeding goals in crop improvement. Computer simulation can be an important resource for the breeder as it provides a means for evaluating dynamic model(s) *in silico* (*e.g.* the genome model, the seed product development 'process' model). Furthermore, it is particularly useful for comparing the process efficiencies without going through field experimentation, thus saving both time and field

resources (Wang et al. 2003). Cases in point: Ribaut et al. (2002) and Ishii and Yonezawa (2007) explored optimal ways to stage the donor crosses to pyramid multiple transgenic events in elite cultivars; Longin et al. (2006) and Gordillo and Geiger (2008a) optimized resources for breeding and testing based on different breeding scenarios; and Gordillo and Geiger (2008a) assessed the reduction in genetic diversity due to selection with a given selection scheme.

Falconer and Mackay (1996) pointed out that quantitative genetic principles underlying particular breeding methods are usually associated with some strong assumptions, some of which may not be met in reality. With computer simulation, we can certainly relax some assumptions and investigate the effects and implications of doing so. Computer simulation can also be used to estimate the effect of population structure. For example, Yu et al. (2008) included the mean of each subpopulation along with the marker effects in the model to interpret the effect of population structure in the maize nested association mapping (NAM) population. Computer simulation is also useful in finding the optimal molecular marker density for marker-based applications like marker-assisted selection (MAS) or predicting performance (Frisch et al. 2000).

As Xu and Crouch (2008) pointed out, modeling and computer simulation are becoming more and more essential for dissecting the genetic basis of complex traits and optimizing breeding approaches. At the same time, the advent of powerful computers has greatly facilitated the assessment and use of different statistical methods for analysis of massive data sets. Advancements in models and methods in computer simulation have been realized; and a number of computer simulation programs have been developed and made publicly or commercially available.

The objective of the paper is to highlight the role of computer simulation in crop genetic improvement, the basics of model building, statistical considerations, and key issues to be

addressed. In addition, we describe the features, functionalities and underlying assumptions for several functionally diverse software packages used *in silico* for the purpose of promoting the use of such tools by plant breeders. These software packages include: Plabsoft, QU-GENE/ QuLine, MBP, GREGOR, PLABSIM, GENEFLOW and COGENFITO (Tables 1.1 and 1.2). More specifically, GREGOR (Tinker and Mather 1993) predicts the average outcome of mating or selection in plant breeding; PLABSIM (Frisch et al. 2000) is a program specifically for simulation of marker-assisted backcross procedures; Plabsoft (Maurer et al. 2008) and GENEFLOW (http://www.geneflowinc.com) focus on certain factors such as genetic diversity; QU-GENE (Podlich and Cooper 1998) uses a specific E(NK) model to estimate the gene-by-environment and epistatic effects; COGENFITO (Hessel et al. 2010) performs as a database searching tools to find specific genotypes; and MBP (Gordillo and Geiger 2008b) optimizes the allocation of testing resources to maximize the genetic gain in maize hybrid breeding. Due to the diversity in the functionality of these programs, a head-to-head comparison of program efficiency and outcome is not possible; however, the models utilized in each program are detailed to the greatest extent possible.

In addition to the computer programs, we also highlight some specific models utilized in predictive breeding simulations (Table 1.3) that may or may not be incorporated in the computer programs reviewed, namely, models for predicting breeding values using genomic selection proposed by Meuwissen et al., (2001) and by Bernardo et al., (2007), and associated algorithms. By examining the basic properties of programs and models and assessing their utility in guiding critical decisions facing breeders, we also intend to suggest possibilities for further development and improvement.

Table 1.1 List of computer software programs examined in this review.

Software	Script language	Platform	Availability (Email or URL)	Reference
Plabsoft	C, R	UNIX, Linux, MS-Win	melchinger@uni-hohenheim.de	(Maurer et al. 2008)
QU-GENE/QuLine	Fortran	MS-Win	http://www.uq.edu.au/lcafs/index.html?page=59974	(Podlich and Cooper 1998)
MBP	C++	MS-Win	andres.gordillo@agreliantgenetics.com	(Gordillo and Geiger 2008b)
GREGOR	Pascal	MS-DOS	http://gnomad.agr.ca/software/gregor/	(Tinker and Mather 1993)
PLABSIM	C++	UNIX, Linux, MS-Win	melchinger@uni-hohenheim.de	(Frisch et al. 2000)
GENEFLOW	unknown	MS-Win	http://www.geneflowinc.com	(Coburn et al. 2002)
COGENFITO	PHP 4.0 with SQL	Web-based	http://maizegdb.org/Cogenfito.php	(Hessel et al. 2010)

This list is not fully inclusive. With the focus on support for decision making in early planning and implementation phases, some related software packages were not taken into account. For example, software for creating linkage maps and for facilitating selection among progeny were not included in the review. In some cases, these have been covered by other reviews (Kearsey and Farquhar 1998). (For a comprehensive listing of computer software programs for genetic analysis, the reader is referred to http://linkage.rockefeller.edu/soft/.) The programs included in this review are those primarily focused on decisions about choice of parents and breeding strategies facing the breeder as he/she devises a plan to meet a defined breeding goal.

1.1.2 Computer simulation programs in review

Plabsoft

Plabsoft is a computer program for population genetic data analyses and simulations in plant breeding under various mating systems and selection regimes. Data analysis routines are provided to analyze both *in silico* and experimental datasets.

Table 1.2 Attributes of computer software programs to guide critical planning decisions in crop improvement: models (or modules), functionalities, and assumptions.

Software	Models/Modules	Summary of Functionality	Assumptions
Plabsoft	Quantitative genetic model[†]; count location model	Integrates population genetic analyses and quantitative genetic models for estimating genetic diversity; tests HWE and calculates LD; haplotype block-finding algorithms to predict hybrid performance	Absence of selection in the base population; random mating; Infinite population size; no crossover interference
QU-GENE /QuLine	E(NK) model[‡]; Infinitesimal model[§]	Employs simple to complex genetic models to mimic inbred breeding programs, including conventional selection and marker-assisted selection	No mutation; no crossover interference; all random terms are normally distributed
MBP	Quantitative genetic model for optimization; Infinitesimal model	Optimizes hybrid maize breeding schemes based on DH lines and maximizes the expected genetic gain per year by means of quantitative genetic model calculations under the restriction of a given annual budget	Timely staggered breeding cycles; no epistatic and maternal effects; no correlated response in testcross performance; infinite population size to calculate selection intensity
GREGOR	Quantitative genetic model	Predicts the average outcome of mating or selection under specific assumptions about gene action, linkage, or allele frequency	No crossover interference; no epistatic effect
PLABSIM	Random-walk algorithm to simulate crossovers during meiosis	Simulates marker-assisted introgression of one or two target genes using backcrossing	No crossover interference
GENEFLOW	Genotype; Pedigree; Population and Report modules; optional Multiplex and Germplasm Security modules	Studies the nature and structure of genetic diversity	Diploid inheritance
COGENFITO	Genome model limited to marker maps in MaizeGDB	Screens marker data from a given genetic mapping population to identify lines with user-defined informative haplotypes.	Maize only

[†] With the quantitative genetic model, an estimate of genotypic value is obtained by summing the allelic effects of a subset of loci, and an estimate of phenotypic value is calculated by adding genetic effects with non-genetic effects, such as environment or error.

[‡] In the E(NK) model, E represents the number of different environment types; N represents the total number of genes involved; the average level of epitasis (K) is calculated by the summation of the level of epistasis in each environment type weighed by their respective frequency of occurrence.

[§] The infinitesimal model assumes a very large (effectively infinite) number of loci each with small effect.

Models & Functionality

Plabsoft employs the count-location model proposed by Karlin and Liberman (1978) to simulate recombination during meiosis. The number of crossovers on a chromosome is determined with a Poisson distributed random variable and the position of each crossover on the chromosome is sampled from a uniform distribution. Genotypic value is estimated as the sum of preset allelic effects of a subset of loci, applying the following equation:

$$G = \sum_{S \subseteq N} X_s ,$$

where X_s is a genetic haplotype effect at a subset of loci S and N is set of all loci (Bulmer 1985). Phenotypic value is estimated by combining genetic effects with non-genetic effects, which are, environmental effects and error.

Some functions in Plabsoft focus on understanding population structure: 1) estimation of allelic and genotypic frequencies and identifying population-specific alleles; 2) measurement of genetic relationship through use of modified Roger's distance (Wright 1978), Nei-Li distance (Nei and Li 1979) and Euclidean distance, which form the basis for cluster analysis and dendrograms; 3) estimates of genetic diversity such as coefficient of gene differentiation between different subpopulations (G_{ST}) (Nei 1973); a similar term to Wright's F_{ST} (1965), representing the relative differentiation between subpopulations); 4) application of principal coordinate analysis (PCoA) to analyze genetic relationship among individuals; and 5) testing of Hardy-Weinberg Equilibrium (HWE) via approximation and Fisher's exact test using Monte Carlo methods (Maurer et al. 2007).

Other functions extended from PLABSIM (a forerunner of Plabsoft; see Section V below) focus on breeding application and simulating individual breeding stages (e.g. backcross) and/or entire plant breeding programs. Users can conduct Pearson's chi-square test and Monte-

Carlo-based Fisher's exact test for gametic and genotypic data to measure linkage disequilibrium; calculate linkage disequilibrium measures like D^2, D', r^2, D'_m, and R for association mapping; conduct haplotype-block-finding algorithms (these algorithms have been employed to predict heterosis and hybrid performance in four factorial crossing designs of a hybrid maize breeding program); and determine the optimum design and investigate the estimated selection response in MAS programs. Prigge et al. (2008) compared the accuracy of computer simulations of recurrent parent genome recovery rate with empirical data from an implemented marker-assisted backcross scheme, and found a high degree of closeness between them.

Properties & Assumptions

Plabsoft is written in C and implemented as a library package to the statistical software R (Ihaka and Gentleman 1996). Currently, the software is supported under both Microsoft Windows and Linux operating system. Users operate the program within the R environment. Input files include: the linkage map, marker data, and trait data files, which can be imported from text files or through interfaces provided by the R environment from databases and from the Plabsoft database (Heckenberger et al. 2008). The genetic architecture of a trait and genome parameters (e.g. ploidy level and chromosome number and length) need to be defined prior to the analysis. The output results can be displayed graphically, e.g. as dendrograms, LD plots, principal coordinate plots and graphical genotypes (Maurer et al. 2008). In Plabsoft, no crossover interference during meiosis is assumed.

Comments

Plabsoft is a program to consider for general population genetic data analysis. From a breeding viewpoint, the haplotype-block-finding algorithm makes it less likely to overestimate the quantitative trait locus (QTL) effects and inaccurately predict hybrid performance (Grapes et

al. 2004). Another advantage is that the program can access external databases, facilitating massive data computation. The software and tutorial may be obtained by contacting the author. Although Plabsoft is highly versatile, it could benefit from a more user-friendly interface.

QU-GENE (version 2.2.04) & QuLine (version 2.0)

QU-GENE (Podlich and Cooper 1998) is a computer simulation platform for quantitative analysis of genetic models aimed at evaluating various breeding strategies and selection approaches in light of GEI and epistasis. Because it considers GEI and epistasis, QU-GENE would likely integrate well with a breeding process model and perhaps a business process model, providing extended utility.

Models & Functionality

QU-GENE compartmentalizes the basic GE system (i.e. the defined genetic and environmental information used in the QU-GENE simulation, referred to as GES) and the application modules (e.g. QuLine) for simulation. QU-GENE is based on the E(NK) model, where E represents the number of different environment types; N represents the total number of genes involved; the average level of epitasis (K) is calculated by the summation of the level of epistasis in each environment type weighed by their respective frequency of occurrence (Cooper and Podlich 2002; Cooper et al. 2007). Thus, this genetic model incorporates the GEI effect and epistasis. Cooper et al., (2007) considered E(NK) as complementary to the basic statistical framework suggested by the equation: P = G + E + GE. In QU-GENE, the phenotypic value is modeled by:

$$p_{ijk} = g_i + e_j + (ge)_{ij} + \varepsilon_{ijk},$$

where g_i stands for the genotypic value, e_j is the macro-environmental effect of environment j, $(ge)_{ij}$ is the interaction effect between genotype i and environment j, and ε_{ijk} is the micro-environmental effect (error)

or

$$p_{ij} = g_{ij} + e_i^b + e_{ij}^w,$$

where the environment effect is divided into two parts: one is between-plot error e_i^b estimated from family mean and the within-plot error e_{ij}^w estimated from broad sense heritability, and g_{ij} is the genotypic value calculated from the user-defined gene action. The former part of the environment effect is used to estimate the among-plot variance; the within-plot error is considered to be environmental noise and is based upon user-defined information.

Built on QU-GENE, QuLine (previously called QuCim) is a genetics and breeding simulation tool that can integrate various genes with multiple alleles operating within epistatic networks and differentially interacting with the environment interaction, and predict the outcomes from a specific cross following the application of a real selection scheme (Wang et al. 2003; Wang et al. 2004). It therefore has the potential to provide a bridge between the vast amount of biological data and breeder's queries on optimizing selection gain and efficiency. QuLine has been used to compare two selection strategies (Wang et al. 2003), to study the effects on selection of dominance and epistasis (Wang et al. 2004), to predict cross performance using known gene information (Wang et al. 2005), and to optimize marker assisted backcross selection to efficiently pyramid multiple genes in elite lines (Kuchel et al. 2005; Wang et al. 2007).

Properties & Assumptions

QU-GENE is written in FORTRAN and implemented under OS Microsoft Windows. The latest version 2.2.04 was released in 2008. Data is imported as the QUG-formatted file, which can be opened and edited in a text editor or directly in the software. QU-GENE supports diploid or amphi-diploid genomic structures with two alleles per locus. A simple graphical user interface has been designed for importing data. Users need to define and specify the information for environment, trait, gene, and population data in the QUG file, especially gene information which

is the fundamental part to define GE system, including gene action (e.g. additive, dominance, and epistasis), GEI, pleiotropy, linkage, multiple alleles, and molecular markers. The GES and POP-formatted files output by the QU-GENE engine are used to store the information about the GE system and an initial population, respectively, for further application in the QuLine module.

QuLine is a QU-GENE strategic application module (current version 2.0) developed to simulate the breeding processes for developing advanced inbred lines. It can also be used to investigate the starting population under the defined GE system. QuLine is the only currently available application module for QU-GENE. QuLine can be implemented under both MS-DOS and MS-Windows environments and has a graphical user interface to facilitate simple operations. The input files for QuLine include GES and POP files (described above), as well as QMP file which contains basic parameters, such as numbers of breeding strategies, runs, cycles, and crosses. Users also need to define the breeding strategies and selection information to be used and to specify the final output results in the QMP file.

With the E(NK) model, the following assumptions are made: a large number of loci with small gene effects contributing to the genetic variation (infinitesimal model), no crossover interference, no mutations and all random terms, such as environmental errors are normally distributed as $\sim N(0, \sigma^2)$.

Comments

QU-GENE is a potentially useful tool to explore the non-additive variation and to make predictions from genetic models defined in terms of gene frequencies, gene action type, and specified target populations of environments. Starting with input on known parameters, estimations of epistasis and genotype by environment interaction will be generated within this framework. The above approach could be utilized to leverage the known genetic information and

to compare breeding strategies by using these in association with the estimated genetic parameters generated by the model.

Some additional notes about QU-GENE are: 1) the sample of environments in a multiple environment trial will not well represent the target population environment if the cycle number is small and the environment is sampled randomly, as the expected frequency of occurrence will not match the real environment types frequency; and 2) no provision is made for mutation, autopolyploid species, or use of doubled haploid (DH) lines.

The user manual illustrates program specifics in detail. Although the data format can be complicated when starting from scratch, the software provides some example files which can be used as templates. We found that run time was brief using the sample files provided, suggesting speed in processing small datasets. The running time would change, of course, depending on the model, parameters, and the size of dataset.

MBP (version 1.0)

MBP (version 1.0; (Gordillo and Geiger 2008b) is a software package to optimize hybrid maize breeding procedures based on DH lines. It is designed to maximize the expected genetic gain per year for a given annual budget and a limited relative annual loss of genetic variance. Alternative breeding plans are optimized using model calculations. The software uses default values for the underlying estimates of variance components and genetic correlation coefficients as well as haploid induction parameters and costs of individual breeding steps based on data from collaborating breeding companies; these can be varied by the user according to his/her genetic, technical, and financial resources.

Models & Functionality

The estimated phenotypic variance between testcross progenies is calculated as the sum of the genotypic variance between testcross progenies and the genotype × year, genotype ×

location, genotype × year × location interaction variances, and the error variance. The estimated genotypic variance between testcross progenies (σ_t^2) comprises the variance of the general combining ability (σ_{GCA}^2) and specific combining ability (σ_{SCA}^2; Griffing, (1956) :

$$\sigma_t^2 = \sigma_{GCA}^2 + \sigma_{SCA}^2 / T$$

where *T* is the number of testers and σ_{GCA}^2 and σ_{SCA}^2 are derived from additive and dominance variance estimates, respectively.

The functions of the program are: 1) maximization of the expected annual genetic gain in general combining ability (GCA) under a restricted annual budget and an upper limit for the decay of genetic variance; 2) inclusion of seven alternative breeding schemes, each differing in the genetic structure of the material used for starting a new breeding cycle, the generation in which the haploid induction is applied, and the stage at which the DH lines are evaluated for *per se p*erformance; 3) prediction of the gain from selection based on index composed of grain yield and dry matter content, implementing numerical methods for the calculation of normal integrals for the distribution of genotypic values under one-, two-, and three-stage selection; 4) specification of the number of lines to be finally selected; and 5) optimizations under a restricted relative annual loss of genetic variance defined by $\Delta\sigma_g^2 = 1/(2N_eY)$, where *Y* is the cycle length in years and effective population size (N_e) is predicted by applying formulae for the joint effects of drift and cumulative selection according to Santiago and Caballero (1995).

Properties & Assumptions

MBP is written in programming language C++ and the compiled program can be implemented under OS Microsoft Windows. Within the confines of a maize breeding program involving DH lines, users may arbitrarily specify the quantitative genetic parameters and operational variables in the input files, such as breeding and mating scheme, the index weights,

the restricted budget and the dimensioning (or allocation) restrictions to be used for the optimization procedures. The software will output the results regarding resource allocation, expected gain in optimization criterion, GCA, and the predicted annual loss of genetic variance.

In MBP, absence of epistatic and maternal effects as well as no correlation between *per se* and testcross performance for yield and harvest moisture is assumed. The program user may specify that the breeding material representing one gene pool is subdivided into multiple, timely staggered breeding cycles (as it is usually the case in practical breeding). When calculating N_e, a constant number of random mating parents and random distribution of number of progenies per cross is assumed. An infinitesimal model of gene effects is assumed to interpret the genetic architecture of a trait and an infinite population size is assumed for calculating selection intensity.

Comments

Although specifically developed to optimize hybrid maize breeding procedures using DH lines, MBP has the potential to be extended to other breeding schemes. Gordillo and Geiger (2008a) illustrated an example of using different genetic variance component ratios, budgets, and breeding strategies. Herein, the authors clarified that differences in the estimated genetic gains between the evaluated optimization variants that are statistically significant remain open, since formulae for computing confidence intervals of genetic gain estimates are available for one-stage selection only (Burrows 1975).

GREGOR (version 1.5)

GREGOR (Tinker and Mather 1993) is a simulation program predicting the average outcome of a given cross or selection scheme.

Models & Functionality

With GREGOR, users can choose from three scenarios, which differ in the number of loci, chromosome number, and crossover rate. All genomes are modeled as diploid. The genotype at every locus in the genome is defined for each individual; a population is defined as a group of specific individuals with conditions in common. The positions and allelic effects of a subset of loci within the genome are specified for defining a trait. Markers are determined to be associated with alleles having dominant or codominant effects at associated loci. The genotypic value estimate is calculated as the sum of allelic effects (additive and dominance effects) of a subset of loci; and the phenotypic value estimate is calculated by adding the genotypic value to an environmental effect that is chosen randomly from a normal distribution.

The function of GREGOR is to predict the average outcome of mating or selection under specific assumptions about gene action, linkage, or allele frequency. It can process various types of population structures e.g. DH lines, recombinant inbred lines (RILs), and diverse mating schemes including random mating, chain crosses, full diallels, and backcrossing. The program also allows individuals comprising the population to be selected manually or randomly.

Properties & Assumptions

GREGOR is written in Pascal and implemented under MS-DOS environment. No empirical data are needed, since all inputs including individual, trait and marker data are simulated by the program. The Haldane mapping function (1919) is used to relate map distance to recombination frequency, or vice versa. The program will output a histogram and graphical genotype for visualization and results consisting of population information, trait and marker phenotypes in a population. In terms of interfacing with other programs, GREGOR can create files readable for Mapmaker/Mapmaker QTL. The program assumes no crossover interference and absence of epistasis.

Comments

This program was designed for illustrating basic breeding methods and quantitative genetic theory. The program tutorial is quite thorough. However, the program has not been updated since 1995. Since GREGOR runs only in a MS-DOS environment, and a DOS emulator such as DOSBox (http://www.dosbox.com/) can be used to facilitate operation in Windows OS. It may have its greatest utility as a teaching aid.

PLABSIM

PLABSIM (Frisch et al. 2000) is a software program which simulates backcross introgression with one or two target genes or genetic factors.

Models & Functionality

PLABSIM simulates recombination during meiosis via the random-walk algorithm (Crosby 1973). A crossover event is considered to have taken place only if a number sampled at random from a uniform distribution is no larger than recombination frequency.

Functions of PLABSIM include simulation of marker-assisted introgression of one or two target genes using backcrossing, evaluation of gene frequencies to estimate the recurrent parent genome proportion in backcrossing, calculation of genotype frequencies to estimate homozygosity and heterozygosity and estimation of the required number of marker data in breeding program. The program allows for population manipulation and selection.

Properties & Assumptions

PLABSIM is written in C++ and implemented under multiple operating systems, such as UNIX and Windows NT. The input file specified by users contains information about linkage map, the base population, and a description of the breeding program, such as various breeding designs implemented, etc. The Haldane mapping function (1919) is used to calculate recombination frequency for the random walk algorithm. The program assumes no crossover

interference. PLABSIM is capable of data export, which provides a handy interface with data analysis programs.

Comments

PLABSIM has narrow focus. Furthermore, it has not been maintained or updated since 2000. However, since PLABSIM functionality has been incorporated into Plabsoft, it is available in a newer and better form.

GENEFLOW

GENEFLOW (www.geneflowinc.com) is a commercial software package designed to support breeding programs and genomic investigations.

Models & Functionality

The software features several built-in modules: a pedigree module to support the analysis of genetic and phenotypic data within the context of a pedigree; a genotype module to identify patterns of inheritance and to compare genetic structure across multiple individuals; a population module to analyze structured populations; and a report module containing a number of reports on genotype-phenotype associations and specific choice of markers and/or progeny for use in breeding. There are optional modules for designing multiplexing kit design, and germplasm security.

GENEFLOW uses pedigree, marker and phenotypic trait data as inputs and allows users to integrate, analyze and visualize this information. It supports simple statistical analyses, such as ANOVA, regression analysis, t-tests, and correlation calculation. Coburn et al. (2002) applied the program to design multiplex panels for SSR polymorphism data analysis in the study of genetic diversity among a wide range of cultivated rice germplasm. Malysheva-Otto et al. (2006) used GENEFLOW to calculate allelic richness, gene diversity, the occurrence of unique and rare

alleles, and the frequencies of heterogeneous loci. This software is a tool for data management and display; model building and data analysis capabilities are limited.

Properties & Assumptions

GENEFLOW is appropriate for the analysis of diploids or functional diploids, such as wheat. Sample databases of rice and barley are included within the system.

Comments

GENEFLOW allows the user to define and execute MAS rules, evaluate parental pairs for target genotype analysis, find polymorphism between any pairs of lines, find lines the most/least similar to a reference or ideotype, and identify non-parental alleles. Its strengths are in data management and display.

COGENFITO

COGENFITO, the composite genotype finder tool, is a web-based program designed to facilitate use of molecular marker data to optimize choice of parents in maize breeding for fine mapping purposes (Hessel et al. 2010). It acts as a browser for large genotype data sets, allowing a user to sort and sift through marker data to identify lines with user-defined haplotypes. The tool has a very limited genome model, in keeping with its focused functionality, and does not have analysis capabilities. The only available implementation at present is at http://www.maizegdb.org/ , where it can be used to browse through more than 5,000 stocks in molecular breeding efforts for maize (Lawrence et al. 2007). Based on a multi-locus genotype query for a particular population of lines, COGENFITO identifies and provides stock center links to genetic stocks that best match the desired parameters. Hessel and colleagues (2010) also provide several proposed use cases, including the use of effectively isogenic lines for testing allelic action and for accelerating positional cloning. COGENFITO is available for use at http://maizegdb.org/Cogenfito.php.

1.1.3 Other resources for model building and computer simulation for prediction purposes

In addition to the software programs described above, some significant work in modeling and simulation methodologies has been done in support of breeder decisions, mainly in constructing genetic models and conducting *in silico* evaluation based on rather narrowly defined product development process models. For example, work by Meuwissen and colleagues (Luan et al. 2009; Meuwissen et al. 2001) and Bernardo and Yu (2007) focused on predicting breeding value using marker and phenotypic data to guide choice of parents, the former in animals and the latter in plants. Details on construction of the genome model as well as statistical methodologies for estimating genetic effects can be useful to the breeder who may wish to develop his/her own tools to guide choice of parents and options for breeding methods. A word is given in review of some of the key issues related to prediction: a) the genetic model, especially in light of the ever-growing body of knowledge about genome architecture and function (Holland 2007; Mackay 2001); b) statistical methods for precision and accuracy; and c) approaches for an integrated strategy.

The genetic model of Meuwissen et al. (2001) was structured as follows. A model genome (1000 cM) was divided into 1000 segments with 100 segments per chromosome, and a QTL was assumed to be located at the midpoint of each segment flanked by two markers. The model suggested that the QTL are identified by flanking markers, and there are on average 50 different haplotypes per centiMorgan. A total of 1010 maker loci were included and about 50,000 haplotype effects were estimated.

Bernardo and Yu (2007) specifically simulated the maize genome (1749 cM) which corresponded to the linkage map published by Senior et al. (1996). The 10 maize chromosomes were divided into N_M bins, where N_M was the number of markers, indicating one marker for each

bin. The location of the marker was set to be at the middle of each bin. A number of bi-allelic QTL influencing the trait of interest were assumed to be randomly located across the genome according to a uniform distribution.

The infinitesimal model is often deployed with prediction of quantitative traits. Laurie et al. (2004) pointed out the futility of modifying a single gene to affect an improvement in oil content in corn, a trait that is polygenic, yet additive in nature and highly heritable. Likewise, other agronomic traits of interest in maize like grain yield, grain moisture, flowering time and plant height, are typically represented by an infinitesimal model (Austin et al. 2000; Buckler et al. 2009; Schon et al. 2004).

In contrast to MAS to identify outstanding progeny wherein markers with strong association with the trait of interest are used to "collect" favorable chromosomal segments to improve performance, approaches to prediction may involve a more inclusive accounting of the genome. For example, GS (or GWS) or approaches that account for epistatic interactions may be employed.

GS was first introduced to predict breeding value in animal breeding (Meuwissen et al. 2001). All marker loci are utilized to estimate breeding value of an individual, not just those associated with a particular level of statistical significance. All gene effects are estimated across the genome simultaneously. Then, individuals are ranked according to the magnitude of the sum of gene effect estimates for each.

Meuwissen et al. (2001) used the following statistical model for predicting breeding value:

$$y = \mu l_n + Xg + \varepsilon$$

where *y* is the vector of phenotype; μ is the overall mean, l_n is a vector filled with 1, X is a design matrix for all genotypes, and *g* is the haplotype effects. The gene effects are assumed to fit a gamma distribution. These are captured by the haplotype whose inheritance is tracked through DNA marker genotyping, assuming no recombination within the haplotype. The random non-genetic effect (**ε**) has a normal distribution with a mean of zero and the variance is adjusted according to the heritability. It is assumed that all genotypes can be traced back to the same base population, which is in HWE and/or linkage equilibrium (LE) and all individuals are unrelated in the base population. Absence of epistasis is also assumed. Bernardo and Yu (2007) later applied the same statistical model to maize breeding, but assumed that distribution of gene effects follow geometric series according to Lande and Thompson (1990).

Statistical methodologies for GS have been evolved progressively. Meuwissen et al. (2001) firstly proposed methods like least square (LS), BLUP, Bayes A and Bayes B (Table 1.3). Calus and Veerkamp (2007) added the polygenic term into the model, and found that the polygenic effect will not increase the GS accuracy although it explains variance components better and removed bias. While de los Campos et al.(2009) used the Bayesian least absolute shrinkage and selection operator (LASSO) to fit marker effects in a regression model to predict gene effects. Compared to shrinkage methods such as Bayes A or Bayes B in Meuwissen et al. (2001), LASSO employed double exponential distribution rather than normal distribution as prior for regression coefficients, imposing more shrinkage on the effect close to zero and less shrinkage on the large effect.

GS applications in plants typically consider environmental effects, and treat these as random (Bernardo and Yu 2007). Piepho (2009) applied GS to accommodate GE data using two-stage analysis and fitted the model by restricted maximum likelihood method. Heffner et al.

(2009) suggested that Bayesian methods might be a better method when handling data with limited phenotypic records and increased marker data sets. Furthermore, they suggested that the selection cycle in plant breeding could be shortened by GS. Wong and Bernardo (2008) concluded that use of GS as a predictive tool was superior to both marker-assisted recurrent selection and phenotypic selection and also increased breeding efficiency by reducing development of new lines by two thirds, from 18 years to 6 years. Later, Bernardo (2009) reported that computer simulations involving GS showed that only three years would be required for genetic improvement of an adapted line using an exotic source.

Table 1.3 Models and methods for predicting breeding value via genomic selection, as per (Meuwissen et al. 2001).

Genetic model	Assumptions	Statistical methods	Specificity
Phenotypic values are obtained by adding genetic effects with nongenetic effects. Gene effects fit gamma distribution. Random nongenetic effects fit a normal distribution with mean of zero and variance adjusted to heritability.	All genotypes can be traced back to the same base population. Unrelated individuals in the base population Absence of epistasis and interactions	**Least-squares (LS) estimation**	Fixed haplotype effects; Only significant effects are estimated.
		Best Linear Unbiased Prediction (BLUP)	Random haplotype effects; Same variance for every segment effect.
		BayesA	Random haplotype effects sampled from normal distribution; Different genetic variance for every segment effect.
		BayesB	Random haplotype effects sampled from normal distribution; Different genetic variance for every segment effect; Some segments have no effects with probability *p*

1.1.4 A look forward

Given the benefits and promising future of computer simulation to guide breeding decisions in early planning and implementation phases, the development of user-friendly software programs and modeling methodologies for this purpose deserves greater focus. Most of the software reviewed has limited focus and/or utility for impacting decisions facing the plant breeders, especially with respect to choice of germplasm and breeding strategy options.

Model building and computer simulation can play an important role in decision making, process design, and operational efficiency in the development of improved crop seed products (Figure 1.1). A model is formulated, e.g. for the genome of the crop of interest, based on available facts, scientific principles, and various assumptions. A model is intended as a simplified version of the system it represents, and the model evolves and is refined as scientific advances are made and more information is available. A simulation tool may, in fact, include more than one model or layers of models. For example, a genome model may represent the genetic architecture of the species of interest; a seed development 'process' model may represent the system used to advance promising materials to commercial release; and a 'business' model may represent the seed roll-out and distribution of a new cultivar. At present, we know a great deal about the genetic systems of various crops and we are learning more every day with advancements in DNA sequencing, biochemistry, and genomics, paving the way for more sophisticated genome models. The model(s) are then used to "predict" the best options based on probable outcomes, whether these are optimal pairs of inbred lines to produce progeny having superior performance or top breeding strategies to minimize development time and maximize the rate of genetic gain, potentially saving time and resources. The predictions can then be validated using actual data, which serves to reinforce and refine or redefine the model and simulation mode. With the incorporation of real data, the program takes the form of a tool, which can be directed to optimize an operational process to improve efficiency. At this level too, continual refinements are made as outputs are captured and applied to further tune the model. Operational efficiencies are realized as the tool is applied to the product R&D pipeline, in terms of accelerated speed to market, greater rates of genetic gain, new knowledge, and/or innovation (i.e. creation of a new type of product or process). As an example, improved breeding schemes for

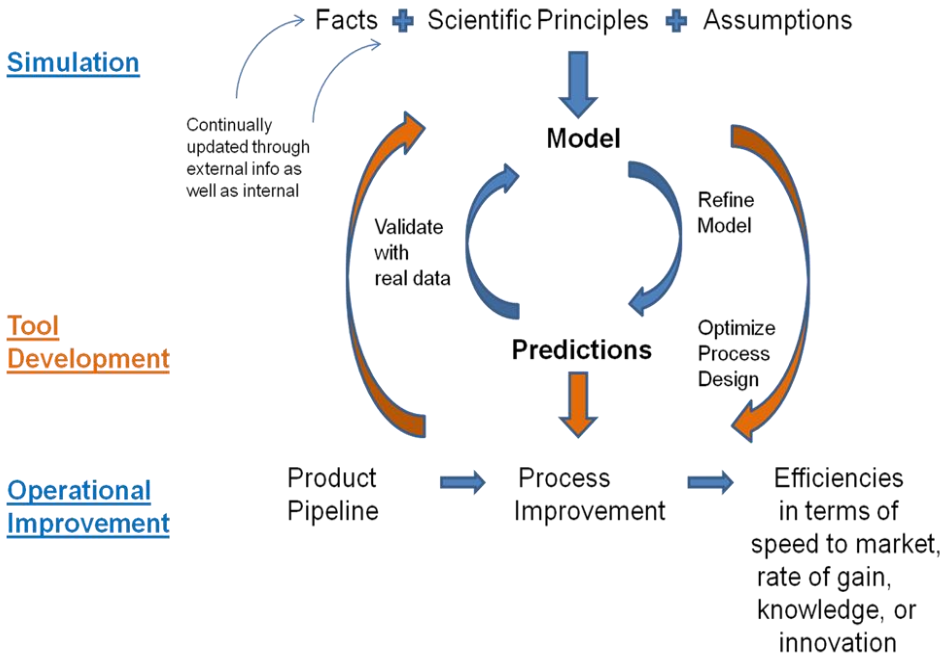
marker-assisted introgression suggested through computer simulation may later be adopted and incorporated in the product pipeline to achieve speed to market in release of corn hybrids with new transgenic trait combinations.

Heffner et al. (2009) suggested that, since speed-to-market and cost efficiency are relatively higher priority in the private sector, software and database development may be given greater attention there than in the public sector. Notwithstanding, Xu and Crouch (2008) emphasized that decision support tools are crucial to analyze and combine other data to make rapid and efficient selection decisions in a short time period while maintaining accuracy. Xu (2010) suggested integrating knowledge from various disciplines in building tools to guide tactical and strategic decisions. Furthermore, such tools are important to the education of the next generation of plant breeders and their preparation for productive careers in crop improvement.

A particular need exist for more support in choosing parents, as decisions in this realm are some of the most crucial facing the breeder, and few current software programs address this area, e.g. MBP and QU-GENE. For instance, when the breeding target is to develop an improved hybrid, computer simulation could be used to model GCA and SCA for predicting the cross performance and optimizing the lines used as parents to create breeding populations (Bernardo 1993, 1994; Piepho 2009). An effective tool would increase the probability of choosing parents that will result in superior new gene combinations, focus resources on materials with the greatest genetic potential, and increase the odds of successfully recovering outstanding progeny. Several issues need to be considered in building models, e.g. investigating the impact of GEI effects, accounting for epistatic effects in the model (Dudley and Johnson 2010), and optimizing marker density to maximize accuracy of prediction. In the maize genome, thousands of genes isolated by repetitive sequence regions contribute small effects to the quantitative traits, in keeping with the

infinitesimal model. This model can be tailored for application to other diploid species, depending on genetic architecture and type of gene action. For example, dominance effects are usually less important in a self-pollinated crop than in a cross-pollinated crop and epistatic effects may be more important in self-pollinated crops (Blanc et al. 2006; Mihaljevic et al. 2005).

Figure 1.1 Computer simulation applied to create tools that improve operational efficiency in seed product development.



Any new simulation program must be able to handle massive amounts of data and offer reduced computing time. The ability to interface with external databases will be critical. New software will likely enable cloud computing including parallel and grid computing, demanding a higher level of design and programming. Furthermore, new simulation tools require testing both *in silico* and experimentally to validate models and methods used.

1.2 Objectives

Given the simulation programs and tools mentioned above, we challenged some of the assumptions commonly incorporated in genome models, namely those of no crossover

interference and invocation of the infinitesimal model of gene effects, with the aim to produce a more accurate representation of the genome. In Chapter 2, we used data from private and public sources to model crossover and distribution of QTL additive and dominance effects. Specifically, a two-pathway crossover interference model (Copenhaver et al. 2002) was used to estimate the strength of crossover interference across the genome based on several doubled haploid data sets contributed by Monsanto Company. Meanwhile, published data from QTL mapping studies were used to derive the distributions of QTL additive effects and dominance coefficients (calculated as the ratio of dominant to additive effect) in the form of mixtures of normals by fitting the Dirichlet Process Gaussian Mixture Model (DPGMM).

Besides absence of crossover interference, epistasis is also typically unaccounted for in most of the genetic simulations and statistical models. For example, Meuwissen et al. (2001) and Bernardo and Yu (2007) fit the data using a linear model to estimate breeding values. However, given the impact of epistasis in the expression of some key traits of interest (Dudley and Johnson 2009), predictive ability may be improved in taking this important genetic characteristic into account. In Chapter 3, we proposed a new nonparametric method, pRKHS, which combines the features of supervised principal component analysis (SPCA) and reproducing kernel Hilbert spaces (RKHS) regression, with versions for traits with no/low epistasis, pRKHS-NE, to high epistasis, pRKHS-E. The new method was evaluated in comparison with current popular methods for practicing genomic selection, specifically RR-BLUP, BayesA, BayesB and RKHS, using both simulated and real data. Beyond prediction, the new method also facilitated inferences about the extent to which epistasis influences trait expression.

In Chapter 4, we applied the genome model mentioned above to a case study involving conversion of a target hybrid for 15 transgenic events. The case study demonstrated how we

could take advantage a more definitive genome model incorporating a more accurate distribution of genetic effects to help to determine the optimal breeding strategy for marker-aided introgression. Prospective breeding strategies differed from each other based on the assignment of various parameters such as: 1) the proportion of non-recurrent parent germplasm remaining in the converted lines; 2) the number of 'versions' of the inbred conversions; and 3) the number of 'versions' of the hybrid conversions needed for a 95% probability of recovering at least one with performance on par with the unconverted target hybrid.

In summary, the following three respective chapters cover: A) Improvement of genome model, including investigation of crossover interference and the distribution of genetic effects; B) Development of a novel statistical method to enhance the predictive ability in choosing parents in the new line development process; and C) Optimization of number of versions of converted hybrid introgressed of 15 transgenic events.

Chapter 2 - Improvement of genome model

2.1 Overview

To serve better predictive ability, the underlying genome model of a simulation tool needs to be synchronized with updated scientific facts and principles. Crossover interference allows us to accurately capture the recombination process within a small genetic interval. At the same time, understanding of the distribution of genetic effects provides a basis for depicting a more realistic picture of genetic architecture, particularly the magnitude and direction of QTL effects across the genome.

Crossover interference has been observed in almost all organisms. However, because not much is known about the degree of interference or the variability among and within individuals, the assumption of no crossover interference is typically invoked in genetic simulation. Taking into account that crossover interference is immensely important when considering the outcome of the recombination process within a small chromosomal interval, this issue must be addressed if we are to accurately predict the results of various breeding strategies focused on eliminating linkage drag in trait introgression. However, whereas crossover interference is a biological process, modeling it is a statistical process. We are employing a two-pathway statistical model to model crossover process across the whole genome and assign genotypes to individuals of the next generation in simulation. The model reflects two possible ways for crossover interference to occur during meiosis: one is called pairing crossover which will not interfere with the crossover occurring adjacently; and the other is called disjunction crossover which will interfere with the crossover occurring adjacently by assigning m gene conversions (non-reciprocal recombination) for the following events before the next crossover occurs. The interference counting parameter (m) and the proportion of pairing crossover (p) across the total number of crossovers are

estimated by using actual data from maize dihaploid populations (contributed by Monsanto Company).

Besides crossover interference, we explored the distribution of genetic effects, specifically additive effects and dominance coefficients, to improve the genome model and provide a stronger basis for accurate simulation of QTL effects. We utilized a Gaussian Mixture Model (GMM) to fit the distribution of genetic effects, which is based on the notion that various alleles fall into a number of classes with different effects. QTL effects collected from previous QTL mapping studies were utilized in this work; however, because the effects reported were observed with additional experimental error and the actual number of mixture components underlying is unknown, a traditional GMM is not sufficient for the problem. We, therefore, derived a Dirichlet Process Gaussian Mixture Model (DPGMM) to include the experimental error into the model and treat the number of mixture components as random variables with a Dirichlet process prior distribution.

2.2 Investigation of crossover interference in maize

2.2.1 Introduction

Meiosis is the specific cell division that is necessary for sexual reproduction, with reproductive cells to be haploid gametes. Prophase I in meiosis has been given special attention mainly in the field of studying recombination and crossover. The DNA exchange events initiated by DNA double-strand breaks could be either reciprocal or nonreciprocal, leading to crossover or noncrossover events, respectively. The distribution of crossover is regarded as a regulated process. For example, regions around centromeres have less frequent recombination events than other regions. Barchi et al. (2008) suggested that crossovers on the same chromosome generally distribute uniformly and broadly whereas Timmermans et al. (1996) indicated that crossover in

maize was restricted to unmethylated and nonrepetitive genomic regions. The different types of distribution of crossover indicated the phenomenon of crossover interference which has been observed in almost all organisms studied (Sturtevant 1915). The distribution of crossovers along a genetic interval is typically inferred from genetic experiments, *i.e.* recombination information, due to the facts that crossovers are not easily observed. Other than indirect inference from recombination information, direct observations from cytogenetic experiments such as late recombination nodules could also be used to indicate the position of crossover (Anderson et al. 2003).

Modeling crossover interference has been not only a biological but also a statistical process. McPeek and Speed (1995) reviewed some statistical models to fit the distribution of crossovers, with certain emphasis on the stationary renewal process model, such as gamma model. Generally, the gamma model could also be named as gamma inter-arrival process, which suggested that the distances between two crossovers are independent and identically distributed from a gamma distribution. The chi-square model is a special case of gamma model with the shape parameter to be integer, *i.e.* $m + 1$, where m is the counting parameter (Zhao et al. 1995). Foss et al. (1993) supported a chi-square model using observed gene conversions from experiments and gave the model some biological meaning, *i.e.* assigning the ratio between noncrossover and crossover to a counting parameter. Copenhaver et al. (2002) later proposed a two-pathway model on the basis of the chi-square model, which suggested the coexistence of two types of crossovers during meiosis: pairing (non-interfering) crossover which doesn't interfere with the crossover occurring adjacently; and disjunction (interfering) crossover which hinders the neighboring crossover by inserting certain noncrossover events between two crossovers. Besides a counting parameter, another parameter, p, was used to govern the

proportion of non-interfering crossovers in the two-pathway model. Later Mézard et al. (2007) extended the two-pathway idea to all plants. And Falque et al. (2009) concluded that two types of crossovers exist in maize genome.

It is essential to find the positions of crossovers for studying crossover interference by either indirect inference from recombination or direct observations from cytogenetic studies. Depending on the organisms, recombination would usually be recorded in the form of two types, *i.e.* single spore data and tetrad data. For example, in Drosophila, a single meiotic product could be retrieved and observed individually as a "single spore", while in yeast and Arabidopsis, products of all four meiosis were regained and observed together as a "tetrad" (Zhao et al. 1995). In maize, it is uncommon and difficult to directly genotype either pollen or eggs to recover a single meiotic product. However, it could be indirectly accomplished by the usage of doubled haploid (DH) technology, an important approach to accelerate inbreeding in plant breeding. With DH, the haploid gamete is steadily reserved in a homozygous diploid status by the doubling chromosome process.

In this study, we employed six DH populations developed and genotyped by Monsanto Company to explore crossover interference in maize. Coverage of 8 chromosomes provided by these data and the two-pathway crossover interference model (Copenhaver et al. 2002) were used to investigate the crossover process. The results obtained from this study were also compared to those in Falque et al. (2009) who fit the late recombination nodules data to a gamma model (McPeek and Speed 1995) considering additional "sprinkle" process (*i.e.* two-pathway interference).

2.2.2 Material and methods

Data sources and processing

Six DH populations derived from F1s were obtained from Monsanto Company to investigate crossover interference in maize. Population sizes ranged from 87 to 148 with each population having different marker coverage (Table 2.1). A total of eight chromosomes were exposed, with different genetic length coverage (Table 2.2). Given parents genotypes, marker data were then converted to recombination scores for fitting the statistical model. For example, we denoted allele originated from parent 1 as 'A' and the one from parent 2 as 'B', and recombination score was recorded as 1 in an interval if the flanking marker genotypes were 'AB' or 'BA'; otherwise 0, indicating no recombination.

Table 2.1 The population size and number of markers of six double haploid populations.

Populations	Pop size	Total # of Markers
1	132	6
2	132	19
3	148	13
4	134	21
5	87	17
6	110	27

Statistical model

The two-pathway crossover interference model adopted from Copenhaver et al. (2002) was employed in this study to assess the crossover interference in maize. The first pathway was designated as the interference-alone pathway which was modeled using the chi-square model (Zhao et al. 1995), denoted as $C_x(C_o)^m$. In detail, due to the fact that crossover inference is a property of meiosis on the tetrad, exchanges events (crossover intermediates) were randomly positioned on the bivalent (four-strand bundle) in light of a Poisson process and were later thinned to get the process on a single meiotic product, *i.e.* pollen or egg. Given the assumption of

no chromatid interference (NCI), the probability a specific crossover goes on a specific single meiotic product is ½ and is independent of the results of other meiotic products. Every exchange event will resolve into either a crossover, denoted as C_x or a non-crossover, denoted as C_o which was biologically interpreted as gene conversion by Foss et al. (1993). According to the $C_x(C_o)^m$ model, every $(m+1)^{st}$ event is a crossover. The interference counting parameter (m) is a non-negative integer and controls the crossover interference strength, with $m = 0$ equivalent to no crossover interference. The other pathway represents the non-interference mechanism, in which case every exchange event will resolve into a crossover. The proportion of non-interfering crossover was governed by additional parameter, p.

Parameters m and p were estimated according to (Copenhaver et al. 2002; Zhao et al. 1995). In detail, a D matrix with dimensions $(m+1)$ x $(m+1)$ was built by assigning a value to the (i, j) element as

$$\sum_{l=0}^{k} \frac{e^{-y} y^{n}}{n!} \binom{n}{l} \left(\frac{p}{p+(m+1)(1-p)} \right)^{l} \left(\frac{(m+1)(1-p)}{p+(m+1)(1-p)} \right)^{n-l} \delta_{(l<k \text{ or } j \geq i)},$$

where $n = (j-i)+k+m(k-l)$ was the total exchange events in that interval, k is the total number of crossover and l is the number of non-interfering crossover, and $y = 2\left(p+(1-p)(m+1)\right)X$ was the DNA double strand breaking length adjusted by exchange events rate, where X is the inter-marker distance provided by Monsanto Company. $\delta_{(l<k \text{ or } j \geq i)}$ is the indicator function and has value of 1 if $l < k$ or $j \geq i$, otherwise 0. According to the theory from Mather (1935) that under the assumption of NCI, if at least one crossover occurs between two markers, the probability that recombination occurs between the two markers on any single meiotic product is ½, matrices N (no recombination) and R (recombination) on j^{th} inter-marker interval were built upon the D matrix to construct the likelihood (Zhao et al. 1995):

Table 2.2 Chromosome coverage and number of markers by six double haploid populations.

Chromosomes	Populations	# Markers	Genetic interval (cM)
Chr 1	1	2	102 - 104
	2	2	55 - 59
		3	253 - 255
	3	2	1 - 1.5
		7	167 - 247
	4	5	1 - 59
	5	2	1 - 2
		2	119 - 122
	6	7	169 - 247
Chr 2	6	2	107 - 113
Chr 3	2	2	103 - 107
		4	162 - 188
	4	4	103 - 129
	5	2	103 - 107
		5	162 - 219
	6	2	162 - 167
		2	218 - 219
Chr 5	2	2	11 - 13
		2	121 - 125
	4	3	104 - 126
	6	4	81 - 107
		2	160 - 172
Chr 6	4	4	99 - 133
	6	4	78 - 103
Chr7	4	2	172 - 174
	5	3	35 - 52
	6	2	66 - 91
Chr 9	2	4	65 - 117
	4	3	65 - 71
	5	3	34 - 67
	6	2	34 - 71
Chr 10	1	4	50 - 62
	3	4	44 - 98

$$N_j = D_0\left(y_j\right) + 1/2\sum_{s\geq 1} D_s\left(y_j\right)$$

$$R_j = 1/2\sum_{s\geq 1} D_s\left(y_j\right),$$

where *s* represents the number of crossover. The likelihood function of observing recombination pattern $\left(i_1, i_2, \ldots i_j, \ldots, i_n\right)$ is

$$L\left(i_1, i_2, \ldots i_j, \ldots, i_n \mid m, p\right) = \frac{1}{m+1}\mathbf{1}M_1 M_2 \ldots M_j \ldots M_n \mathbf{1}',$$

where i_j is the recombination score in j^{th} interval, which has two possible values: 0 for no recombination and 1 for recombination in j^{th} interval and M_j was the matrix dependent on the recombination score, *i.e.* $M_j = N_j$ when $i_j = 0$, and $M_j = R_j$ when $i_j = 1$. A grid search using optimization function "*optim*" in R (R Development Core Team 2011) was used to estimate *m* and *p* that minimize the negative log-likelihood. Parameter *m* was restricted to take integers from 0 to 20 and variable *p* was in the boundary of [0, 1](Copenhaver et al. 2002). Method "L-BFGS-B" (Byrd et al. 1995) was used in "*optim*" function due to its "box constrains" feature. Both chi-square model (null model with only parameter *m*) and two-pathway models (alternative model with parameter *m* and *p*) were used to fit the data and the log-likelihood ratio test was used to determine which model fit the best. The distribution of the test statistic follows χ_1^2.

2.2.3 Results

Chromosome 1 was covered in all six populations and thus was the one that had the longest chromosome coverage (Table 2.3). Most of the *p* estimators were zero, suggesting interference-alone model works well for the linkage groups (LKG) in chromosome 1. Out of nine LKGs in chromosome 1, four of them showed no crossover interference, *i.e.* *m* = 0, three showed small to moderate crossover interference ($1 \le m \le 10$), and the other two fragments showed strong crossover interference, *i.e.* $15 \le m \le 20$. Two LKGs from population 3 and 6 were mostly overlapped in the genetic interval (167 – 247 cM) but had different outcomes. Interference of LKG from population 3 had higher strength than that from population 6. Similar to chromosome 1, chi-square model worked well in the 5 cM region for chromosome 2 and a high level of interference was detected. On chromosome 3, no or low crossover interference was found (Table

2.3). Out of seven LKGs, four were determined as no crossover interference, two were in low interference, and only one LKG spanning 21 cM was fitted by high *m* value. Among the four LKGs without crossover interference, two were overlapped at regions from 103 to 107cM. The LKG ranging from 162 to 219 cM was explained by null model as well as the two-pathway model.

No two-pathway crossover interference was detected in LKGs of chromosome 5, 6, and 7 (Table 2.4). On chromosome 5, chi-square model fit the data with various strength of interference in different regions. Out of five LKGs, three showed no crossover interference and two were detected as strong crossover interference ($15 \leq m \leq 20$). Regions on chromosome 6 were demonstrated in two populations and showed moderate (m = 5) to strong (m = 20) crossover interference. On chromosome 7, three non-overlapped LKGs provided by three different populations exhibited no, moderate strong and strong crossover interference.

On chromosome 9, LKG from population 2 displayed overall no crossover interference in the region of 65 to 117 cM, part of which (65 – 71 cM) showed strong crossover interference. LKGs from population 5 and 6 shared most of the genetic interval and were explained by two-pathway model as well as the chi-square model. In the region (34 – 67 cM), strong interference was detected by two-pathway model with 17% of the crossovers to be non-interfering while low interference (m = 2) was detected by chi-square model. The LKG (34 – 71 cM) was detected with low crossover interference using both models. On chromosome 10, the LKG spanning 12cM was detected with strong interference while the LKG ranging from 44 to 98 cM was detected with low interference (Table 2.4).

Table 2.3 Estimates of *m* and *p* (only in alternative model, *i.e.* two-path way model) on chromosomes 1, 2 and 3.

Chromosomes	Populations	$\hat{m}$	$\hat{p}$[a]	Genetic Interval (cM)	Log-likelihood ratio test statistic	*p*-value[b]
Chr 1	1	0 0	0.000	102 - 104		
	2	20 20	0.000	55 - 59		
		0 0	0.000	253 - 255		
	3	8 8	0.000	1 - 1.5		
		5 5	0.006	167 - 247	0.001	0.97
	4	0 0	0.000	1 - 59		
	5	0 0	0.000	1 - 2		
		17 17	0.000	119 - 122		
	6	1 1	0.000	169 - 247		
Chr 2	6	20 20	0.000	107 - 113		
Chr 3	2	0 0	0.000	103 - 107		
		2 2	0.000	162 - 188		
	4	20 20	0.000	103 - 124		
	5	0 0	0.000	103 - 107		
		1 1	0.420	162 - 219	0.213	0.64
	6	0 0	0.000	162 - 167		
		0 0	0.000	218 - 219		

[a] The null model (chi-square model) doesn't have estimates for *p*.

[b] *p*-value is shown only when parameter *p* has non-zero value in the alternative model.

Table 2.4 Estimates of *m* and *p* (only in alternative model, *i.e.* two-path way model) on chromosomes 5, 6, 7, 9 and 10.

Chromosomes	Populations	$\hat{m}$	$\hat{p}$[a]	Genetic Interval (cM)	Log-likelihood ratio test statistic	*p*-value[b]
Chr 5	2	0 0	0.000	11 - 13		
		16 16	0.000	121 - 125		
	4	0 0	0.000	104 - 126		
	6	20 20	0.000	81 - 107		
		0 0	0.000	160 - 172		
Chr 6	4	5 5	0.000	99 - 133		
	6	20 20	0.000	78 - 103		
Chr 7	4	13 13	0.000	172 - 174		
	5	20 20	0.000	35 - 52		
	6	0 0	0.000	66 - 91		
Chr 9	2	0 0	0.000	65 - 117		
	4	20 20	0.000	65 - 71		
	5	20 2	0.174	34 - 67	0.244	0.62
	6	1 1	0.346	34 - 71	0.214	0.64
Chr 10	1	20 20	0.000	50 - 62		
	3	1 1	0.369	44 - 98	0.538	0.46

[a] The null model (chi-square model) doesn't have estimates for *p*.

[b] *p*-value is shown only when parameter *p* has non-zero value in the alternative model.

2.2.4 Discussion

Overall, interference-alone model (chi-square model) works well for the six datasets, indicating single pathway exist in the covered maize linkage groups. However, the results that 13 out of 33 observed LKGs were detected to have no crossover interference (Table 2.3 and 2.4), demonstrated that across the whole genome some genetic regions do not experience interference which conformed to the conclusions made by Falque et al. (2009). Falque et al. (2009) suggested the coexistence of two pathways of crossover in maize genome and identified that the crossover interference strength was moderate across the maize genome with strength parameter *v* being around 6 to 8 (*i.e.* $m = v - 1$ is around 5 to 7) and the proportion of noninterfering crossovers had a range from 6 to 23%.

We attributed the deviation of our results from Falque et al. (2009)'s work to two reasons. First of all, due to different data source, variations of the results were expected. The data used in this study was recombination data which was used to indirectly infer the distribution of crossover, while Falque et al. (2009) used the direct observation of late recombination nodule to indicate positions of crossover. The other reason may be due to the fact that most of the recombination data used in this study spanned small genetic intervals, *i.e.* we explored the interference in various small blocks, while Falque et al. (2009) investigated the interference using information from the whole chromosomes. To deal with the above issue that chromosomes were not fully covered by our data, we also applied a regression based method proposed by Housworth and Stahl (2009) to estimate parameter *p* on a chromosome-wide basis and found relatively small *p* estimator overall, indicating most of the crossovers are in interfering pathways across the genome (data not shown). Meanwhile, the results that large estimates of interference strength parameter *m* were detected in the present study conformed to the suggestion from

Copenhaver et al. (2002) that within the frame of interference-alone model, green plants have high crossover interference.

It is noticed that the estimates of counting parameter *m* are detected with large variations in certain genetic intervals. One possible explanation may be that estimator *m* has large standard errors, which is supported by the results obtained through simulations that given true parameter, the estimated *m* could be either large or small within certain boundary (Copenhaver et al. 2002). The confidence intervals of interference parameters are hard to obtain analytically because we only explore *m* from 0 to 20 in the study for the ease of computation, leading to a skewed distribution of *m*. Moreover, the issue of small population sizes may also cause the problem of failure to obtain certain recombination events which further leads to inaccurate estimates.

Given the estimates obtained from various linkage groups, the final estimates of *m* could be obtained by taking the weighted mean of the estimates. In detail, we would split the chromosome into several bins, *e.g.* 10, with each bin spanning certain genetic distances and the weight of the linkage group was thus calculated as the percent of genetic interval coverage. And *m* is calculated as $m = \sum_{i=1}^{l} w_i * m_i \Big/ \sum_{i=1}^{l} w_i$, where w_i is the weight and m_i is the *m* estimator of i^{th} bin.

The information obtained from the study could guide simulation of different population structures, such as backcross, RIL, DH and F_2, etc. The gametes simulated using crossover interference model may better represent the true recombination patterns (Zhao et al. 1995). The simulated genetic populations could be further employed to construct various breeding programs, such as genomic selection, and marker assisted backcross, etc. Among those, application of crossover interference on marker assisted backcross (MABC) is promising due to the fact that linkage drag in MABC could be a huge problem especially when there are deleterious genes in

the flanking regions of the transgenic event. And one of the problems of eliminating linkage drag is to evaluate if there are sufficient crossovers to switch the foreign DNA out, in other words, if the crossover interference is high in the flanking regions then we need larger population size to have a crossover occur between the event and foreign DNA.

2.3 Investigation of distribution of genetic effects in grain crops

2.3.1 Introduction

Numerous QTL mapping studies have been reported with various grain crops, e.g. rice, wheat, and maize, etc. in the last decades (*e.g.* Briggs et al. 2007; Sun et al. 2010; Wang et al. 2010). However, very few publications have explored the distribution of QTL effects in crops. Typically, the infinitesimal model, which assumes an infinite number of loci of small effect, is invoked to explain the observed genetic variation because it can easily be incorporated into the statistical analysis, especially for best linear unbiased predication of breeding values (Henderson 1984). However, there is mounting evidence that most quantitative traits are actually controlled by a few loci with large effects and a large number of loci with small effects (Bennewitz and Meuwissen 2010; Bost et al. 2001; Bost et al. 1999; Hayes and Goddard 2001). Under the latter hypothesis, distributions of various types have been proposed to represent genetic effects, i.e., negative exponential distribution (Otto and Jones 2000; Xu 2003a), gamma distribution (Hayes and Goddard 2001), inverse chi-squared prior distribution (Meuwissen et al. 2001), and geometric series (Lande and Thompson 1990). Recently, Bennewitz and Meuwissen (2010) applied the results from three pig F_2 mapping populations evaluated for meat quality and carcass traits to infer the distribution of additive effects and dominance coefficients, fitting a Gaussian Mixture Model (GMM). The idea of utilizing GMM is based on the notion that various alleles fall into a number of classes with different effects (Bennewitz and Meuwissen 2010). The merit

of employing GMM is its flexibility with different combinations of mixtures of normals leading to different shapes of the distribution.

In the finite mixture model, the number of components *K* is pre-specified. The value could be determined based on some specific information or criteria, such as Akaike information criterion (AIC) and Bayesian information criterion (BIC). This requirement frequently encountered in parametric statistics could be sidestepped by introducing a nonparametric Bayes, Dirichlet process, which assumes an infinite number of components. The Dirichlet process is defined as a random process by which a sample drawn is a random discrete distribution; it can be considered a 'distribution over distributions' and has been used widely in the field of population genetics to explain population structure (Gao et al. 2007; Huelsenbeck and Andolfatto 2007).

Modeling the distribution of QTL additive and dominant effects can lead to greater insights into the underlying genetic characteristics of quantitative traits and assisting plant breeding in several areas. That is, more accurate representation of QTL effects would reduce bias and, in turn, lead to more reliable genetic simulation of breeding strategies. In addition, applied to choice of parents or selection among progeny, estimates of breeding value may be improved by using more realistic estimates of QTL effects for assignment of priors. For example, genomic selection based on Bayesian models currently assumes the variance of additive effects to be sampled from an inverse chi-squared prior distribution (Meuwissen et al. 2001), which was further thought to be influenced by the arbitrary selection of hyperparameters (Gianola 2009). In this study, we applied a particular case of the Dirichlet process mixture model, namely Dirichlet process Gaussian mixture model (DPGMM), to derive the distribution of QTL effects in various grain crops, specifically maize, rice and wheat. The additive QTL effects represented results from four different mapping studies using recombinant inbred lines (RIL) or backcross

populations across the three crops evaluated for yield, agronomics, morphological and domestication, and grain quality traits. QTL dominant coefficients represented results from a meta-analysis using five maize QTL mapping studies with five F_2 or F_3 crosses in corn evaluated for yield, stress tolerance, morphological, and grain/silage quality traits. Two limitations of the data, namely that the observed QTL effects were estimated with errors and that only statistically significant QTL effects were reported, were taken into account.

2.3.2 Materials and methods

Data sources

Additive QTL effects were assembled from previous QTL mapping studies performed in corn (Briggs et al. 2007; Messmer et al. 2009), rice (Wang et al. 2011), and wheat (Sun et al. 2010); see Table 1 for a list of traits and number of QTL from each data set. Composite interval mapping had been used to map quantitative trait loci in three out of four studies, while multiple interval mapping had been applied by Briggs et al. (2007). In corn, Messmer et al. (2009) had evaluated a recombinant inbred line (RIL) population derived from a cross between two subtropical white dent maize lines to map genes controlling yield components and secondary traits in corn, whereas Briggs et al. (2007) had utilized a maize-teosinte backcross (BC_1) population to explore genes controlling domestication and morphological traits such as plant architecture, primary tassel and lateral inflorescence. A total of 57 and 59 QTL additive effects had been identified in the former and later experiment, referred to as corn data I and corn data II, respectively. Rice data had originated from a QTL study based on RILs evaluated for fourteen agronomic traits including grain weight, plant height, grain length (Wang et al. 2011). Out of 49 mapped QTLs, 45 QTL additive effects were used in this study for further analysis. Wheat data based on 132 RILs evaluated in seven environments (Sun et al. 2010) for kernel weight, test

weight, kernel diameter, grain protein content, and kernel hardness index contributed 44 significant QTL additive effects. The histograms of observed additive effects of four studies were shown in Figure 2.1.

Figure 2.1 Histogram of observed QTL additive effects from a) corn data I, b) corn data II, c) rice data, and d) wheat data. The values of additive effects are in unit of phenotypic standard deviation.

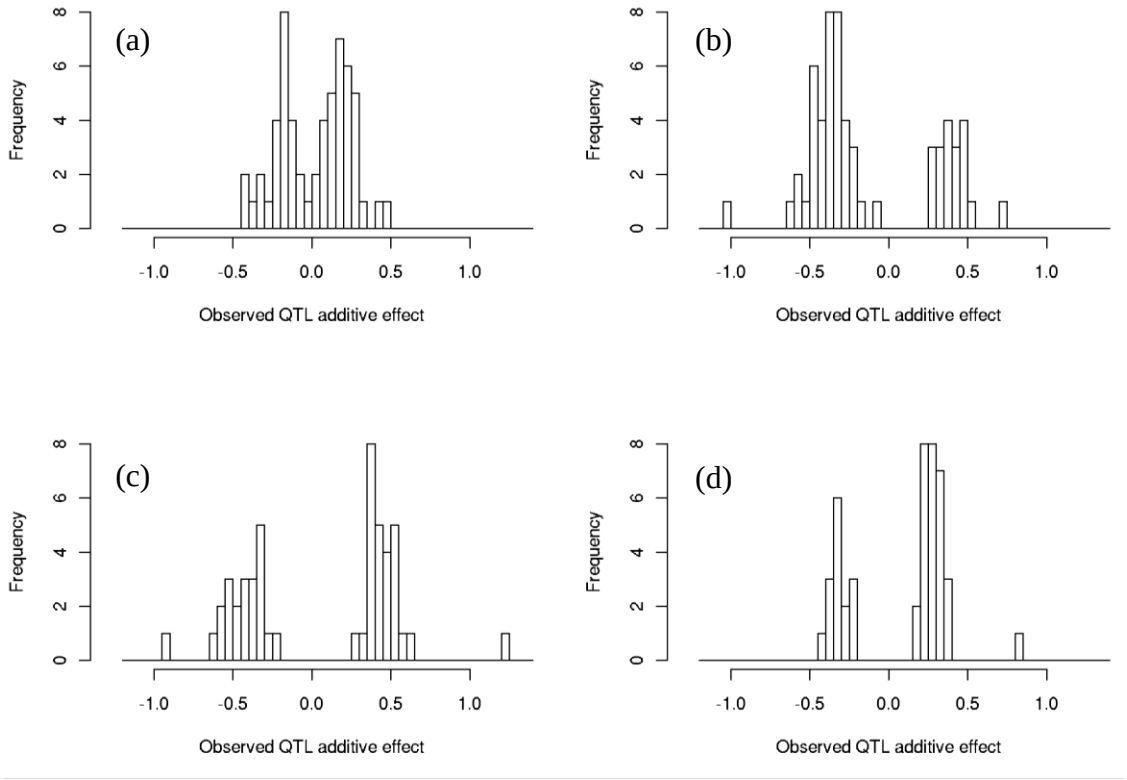


Table 2.5 List of number of QTLs that were associated with traits for studying QTL additive effects.

Data sets	Traits	QTL number
Corn data I	Days to anthesis	12
	Anthesis-to-silking interval	8
	Grain yield	5
	Kernel number	7
	100-kernel weight	11
	Plant height	14
Corn data II	Branch number	2
	Cob diameter (teosinte)	4
	Culm diameter	1
	Cupules per rank	2
	Days to pollen	4
	Glume score	5
	Inflorescence length	2
	Lateral branch internode	3
	Lateral branch	2
	Lateral inflorescence branch	1
	Length of central spike	2
	Male spikelet length	3
	Mean lateral branch internode	2
	Number of barren nodes	1
	Number of tassel branches	5
	Percent staminate spikelets	3
	Plant height (teosinte)	6
	Prolificacy	2
	Ranks of cupules	3
	Tassel branching space length	5
	Tillering	1
Rice data	Heading date	1
	Culm diameter	3
	Plant height	4
	Flag leaf length	3
	Flag leaf width	4
	Tiller angle	3
	Tiller number	1
	Panicle length	5
	Grain length	4
	Grain width	5
	Grain thickness	4
	1,000-grain weight	4
	Spikelet number	4
Wheat data	Test weight	11
	kernel weight	9
	Kernel diameter	10
	Grain protein content	4
	NIR-hardness index	5
	SKCS-hardness index	5

Table 2.6 List of number of QTLs that were associated with traits for studying QTL dominance coefficients.

Populations	Traits	QTL number
Pop1	Kernel oil concentration	11
Pop2	Root angle	10
	Plant height	5
Pop3	*in vitro* dry matter digestibility	4
	in vitro cell wall digestibility	3
	Neutral detergent fibre	4
	Acid detergent fibre	5
	Water-soluble carbohydrate	2
	Kernel oil content	4
	kernel protein content	4
	Kernel starch content	5
Pop4	Stripe virus resistance	6
Pop5	Grain yield	3
	100-kernel weight	9
	Kernel number per ear	6
	Cob weight per ear	7
	Kernel weight per ear	3
	Ear weight	5
	Ear number per plant	5

In addition, dominance coefficients which were defined as the ratio between the observed QTL dominance deviation and absolute value of QTL additive effects were assembled from maize mapping studies. The absolute value of additive QTL effects was used because the sign of QTL effect only indicated which parent had contributed the favorable allele, not the true direction of specific additive effect. Due to limited data published in individual studies, dominance coefficients were extracted from a meta-analysis using five maize QTL mapping studies with F_2 or F_3 populations. A total of 101 significant quantitative trait loci were assembled (Table 2.6), among which 1) 11 QTLs for kernel oil concentration were mapped in an F_2

population (Song et al. 2004); 2) 15 QTLs for root angle and plant height were mapped in an F_2 population (Omori and Mano 2007); 3) 31 QTLs for stalk digestibility and kernel composition in a F_3 population (Wang et al. 2010); 4) 6 QTLs for stripe disease resistance were mapped in an F_2 population (Dintinger et al. 2005); and 5) 38 QTLs for drought tolerance, yield and yield components were mapped in an F_3 population (Xiao et al. 2005). All five mapping studies shared a common parent, B73 and employed composite interval mapping to detect QTLs.

Data processing

The standard error (SE) for additive QTL effects and dominance coefficients was measured to take into account the experimental error. If LOD scores for QTL were absent, standard errors were generated by taking sample standard deviation of effects from multiple environments. The SE of corn data II and data from (Dintinger et al. 2005) were produced in this way, where only those QTLs detected in at least two environments were included in the final dataset. For rest of the studies, SEs were derived from LOD scores as suggested by Hayes and Goddard (2001). Standard errors of dominance coefficients were estimated by the delta method suggested by Bennewitz and Meuwissen (2010), assuming no covariance between additive and dominance effects. In detail, $SE_{d/a} = (d/a) * \sqrt{\left(\frac{SE_a}{a}\right)^2 + \left(\frac{SE_d}{d}\right)^2}$, where SE_a and SE_d were standard errors for additive effects a and dominance effects d, respectively.

Additive QTL effects were scaled by their corresponding phenotypic standard deviations (PSD) in order to combine data across traits. PSD were computed based on the following in order of priority: 1) raw data, 2) error variance and heritability, and 3) range of phenotype values. Phenotypic range was assumed to be 8 PSD, considering that most traits follow a normal distribution. Since for the data sets from which dominance effects were generated none of these conditions were fulfilled to obtain the PSD for additive QTL effects, the additive effects from

those five corn studies were not utilized in analyzing the distribution of additive QTL effects. Note that scaling process was not necessary for dominance coefficients, because PSD was canceled out in the *d/a* ratio.

Due to limited power of QTL mapping studies (Churchill and Doerge 1994), many QTLs with near-zero effects could not be detected, causing a truncation of the additive QTL effects distribution. Faced with this issue, Bennewitz and Meuwissen (2010) suggested a "doubling" process to manage the data, given the assumption that QTL effects occur at the highest frequency around zero. Basically, both signs for the same QTL additive effect were created. For example, for i^{th} effect y_i with SE τ_i, $-y_i$ was added to the data with the same SE. The above procedure ensures the mean of the distribution of additive effects is zero. The "doubling" process was not applied to dominance coefficients, because most loci have observed effects around zero.

Dirichlet Process Gaussian Mixture Model (DPGMM) and priors

We modeled the distribution of additive QTL effects and dominance coefficients using mixtures of normal distributions, namely GMM (Rasmussen 2000). The goal was to assign genetic effects to different mixture components, based on the similarity of their values. Two latent variables were introduced, 1) the total number of mixture components (cluster size, *K*) and 2) the assignment of i^{th} effects to components (cluster indicator, $c_i \in \{1,...,K\}$). The GMM model was modified to expand the experimental error term:

$$p(y_i \mid \lambda_1,\ldots,\lambda_K) \sim \sum_{k=1}^{K} \pi_k N\left(y_i; \mu_k, \sigma_k^2 + \tau_i^2\right), \qquad [1]$$

where y_i was the i^{th} observed QTL effect, τ_i was the known standard error of i^{th} effect, and $\lambda_k = \{\pi_k, \mu_k, \sigma_k^2\}$ was the k^{th} parameter set, where variables π_k, μ_k and σ_k^2 were the mixing

proportion, mean and variance of the k^{th} mixture component, respectively. The GMM could be formulated hierarchically as follows:

$$\begin{aligned} p(y_i \mid c_i, \mathbf{\Lambda}) &\sim N\left(y_i; \mu_{c_i}, \sigma_{c_i}^2 + \tau_i^2\right) \\ c_i \mid \pi_{1:K} &\sim Discrete(\pi_1, \pi_2, \ldots \pi_K) \\ \left(\mu_k, \sigma_k^2\right) &\sim G_0 \\ \pi_1, \pi_2, \ldots \pi_K &\sim Dirichlet(\alpha/K, \ldots, \alpha/K), \end{aligned} \qquad [2]$$

where G_0 was a joint prior distribution for $\left(\mu_k, \sigma_k^2\right)$ and mixing proportions $\pi_{1:K}$ were drawn from a symmetric Dirichlet distribution with α to be concentration parameter. Conditional on the mixing proportions, the latent indicator variables c_i 's were sampled from discrete distribution, specifically multinomial distribution. By integrating out mixing proportions, the prior for the c_i in model [2] could be written as a probability conditional on $\mathbf{c}_{-i}$ (Neal 2000):

$$p(c_i = k \mid \mathbf{c}_{-i}, \alpha) = \frac{n_{-i,k} + \alpha/K}{n - 1 + \alpha},$$

where $n_{-i,k}$ was the number of effects, not including y_i that were linked with class *k*. As *K* goes to infinity, the limits of the prior for the c_i reach the following:

$$p(c_i \mid \mathbf{c}_{-i}, \alpha) = \begin{cases} \dfrac{n_{-i,k}}{n - 1 + \alpha} & c_i = k, n_{-i,k} > 0 \\ \dfrac{\alpha}{n - 1 + \alpha} & \forall i \neq i', c_i \neq c_{i'} \end{cases}. \qquad [3]$$

As $K \to \infty$, the Dirichlet distribution becomes a Dirichlet process (DP) in the limit (Ferguson 1973; Neal 2000). The infinite limit of model [2] thus could be written as a DPGMM:

$$\begin{aligned} y_i &\sim N\left(y_i \mid \theta_i, \sigma_i^2\right) \\ \left(\mu_i, \sigma_i^2\right) &\sim G \\ G &\sim \mathrm{DP}(\alpha, G_0), \end{aligned} \qquad [4]$$

where $\theta_i \sim N\left(\mu_i, \tau_i^2\right)$ was a nuisance parameter, *G* was a random discrete distribution drawn from DP, and G_0 was the base distribution, which specified the joint prior distribution of $\left(\mu_i, \sigma_i^2\right)$. Given that the regular choice of priors for the mean and variance of the Gaussian are Normal and Inverse Gamma distributions, respectively, conjugate joint priors $N\left(\mu_i; \mu_0, \sigma_0^2\right) * IG\left(\sigma_i^2; r_1, r_2\right)$ were chosen in the model.

Gibbs sampling

In Bayesian framework, unknown variables were sampled and updated from the conditional posterior distribution using Markov Chain Monte Carlo (MCMC) (Robert and Casella 2004). By Bayes rule, the joint posterior distribution is proportional to the product of the prior and likelihood. Considered the likelihood and priors in [3] and [4], the full joint posterior distribution was written as follows:

$$\begin{aligned} p\left(\mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 \mid \mathbf{y}\right) &\propto \prod_{i=1}^{n} N\left(y_i; \theta_i, \sigma_i^2\right) \pi\left(\theta_i, \mu_i, \sigma_i^2, c_i\right) \\ &= \prod_{i=1}^{n} N\left(y_i; \theta_i, \sigma_i^2\right) N\left(\theta_i; \mu_i, \tau_i^2\right) N\left(\mu_i; \mu_0, \sigma_0^2\right) IG\left(\sigma_i^2; r_1, r_2\right) p\left(c_i \mid \mathbf{c}_{-i}, \alpha\right). \end{aligned} \quad [5]$$

Unobservables $\left(\mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2\right)$ were repeatedly sampled and updated from their posteriors conditional on all other variables. The Gibbs Sampler was implemented as follows:

1) Initialization. Assign initial values for $\left(\mu_k, \sigma_k^2\right)$ where $k = 1$ and $c_i = 1,$ for $i = 1{:}\, n$.

2) Update θ_i: The conditional posterior distribution of θ_i was

$$P\left(\theta_i \mid else\right) \propto N\left(y_i; \theta_i, \sigma_k^2\right) N\left(\theta_i; \mu_k, \tau_i^2\right) \sim N\left(\theta_i; \frac{\frac{y_i}{\sigma_k^2} + \frac{\mu_k}{\tau_i^2}}{\frac{1}{\sigma_k^2} + \frac{1}{\tau_i^2}}, \frac{1}{\frac{1}{\sigma_k^2} + \frac{1}{\tau_i^2}}\right)$$

3) Update cluster indicators c_i: The conditional posterior probabilities for c_i were:

$$P(c_i = k \mid else) \propto N(y_i; \theta_i, \sigma_k^2) N(\theta_i; \mu_k, \tau_i^2) p(c_i \mid \mathbf{c}_{-i}, \alpha)$$
$$= \frac{n_{-i,k}}{2\pi\sqrt{\sigma_k^2 \tau_i^2}} \exp\left(-\frac{(y_i - \theta_i)^2}{2\sigma_k^2} - \frac{(\theta_i - \mu_k)^2}{2\tau_i^2} \right),$$

$$P(c_i = K+1 \mid else) \propto \alpha \iint N(y_i; \theta_i, \sigma_{k+1}^2) N(\theta_i; \mu_{k+1}, \tau_i^2) N(\mu_{k+1}; \mu_0, \sigma_0^2) IG(\sigma_{k+1}^2; r_1, r_2) d\mu_{k+1} d\sigma_{k+1}^2$$
$$= \frac{\alpha}{2\pi} \frac{r_2^{r_1}}{\Gamma(r_1)} \frac{\Gamma\left(r_1 + \frac{1}{2}\right)}{\left(\frac{1}{2}(y_i - \theta_i)^2 + r_2\right)^{r_1 + \frac{1}{2}}} \sqrt{\frac{1}{(\tau_i^2 + \sigma_0^2)}} \exp\left(-\frac{(\theta_i - \mu_0)^2}{2(\tau_i^2 + \sigma_0^2)} \right),$$

where $\Gamma(.)$ is the gamma function and $k \in \{1 \ldots K\}$. Note that constant $\frac{1}{n-1+\alpha}$ was omitted in both probabilities and $(\mu_{K+1}, \sigma_{K+1}^2)$ were unknown and needed to be integrated out to leave c_i as the only state of Markov Chain. DP was represented via the Chinese Restaurant Process (CRP) (Aldous 1985). Effects were assigned to either currently holding cluster(s) or a new cluster based on the above probabilities. If a new cluster was chosen, then the cluster size was increased, *i.e.* $K+1 \to K$. In case of $n_{-i,k} = 0$, the k^{th} cluster was eliminated and the cluster indicators were decreased by one, *i.e.* $K \to K-1$.

4) Resample and update (μ_k, σ_k^2) suggested by Algorithm 2 as per Neal (2000) as follows:

$$P(\mu_k \mid \theta_i \in k^{th} cluster, else) \propto \prod_{i=1}^{n_k} N(\theta_i; \mu_k, \tau_i^2) N(\mu_k; \mu_0, \sigma_0^2)$$
$$\sim N\left(\mu_k; \frac{\sum_{i=1}^{n_k} \frac{\theta_i}{\tau_i^2} + \frac{\mu_0}{\sigma_0^2}}{\sum_{i=1}^{n_k} \frac{1}{\tau_i^2} + \frac{1}{\sigma_0^2}}, \frac{1}{\sum_{i=1}^{n_k} \frac{1}{\tau_i^2} + \frac{1}{\sigma_0^2}} \right)$$

$$P(\sigma_k^2 \mid y_i \in k^{th} cluster, else) \propto \prod_{i=1}^{n_k} N(y_i; \theta_i, \sigma_k^2) IG(\sigma_k^2; r_1, r_2)$$
$$\sim IG\left(\sigma_k^2; r_1 + \frac{n_k}{2}, \frac{1}{2} \sum_{i=1}^{n_k} (y_i - \theta_i)^2 + r_2 \right),$$

where n_k is the number of effects associated with the k^{th} mixture component.

5) Repeat steps 2 to 4.

Gibbs sampler was implemented with 100,000 iterations to update conditional posterior distributions. The first 80,000 samples were discarded as burn-in and the rest 20,000 samples were used to construct joint posterior distribution. The hyper-parameters in [5] were set to be $\alpha = 0.05,\ r_1 = 1,\ r_2 = 0.01,\ \mu_0 = 0,\ \sigma_0^2 = 0.01.$ Convergence was checked by inspection of negative log-likelihood plots. After burn-in period, when the Markov chain converges to the stationary distribution, sampled parameters were collected to form the posterior distribution. We employed posterior means for estimating $\left(\hat{\mu}_k, \hat{\sigma}_k^2\right)$ and posterior modes for estimating $\hat{c}_i$, which was further used to infer $\hat{\pi}_k$. The Bayesian Confidence Interval (BCI) which was the counterpart of the confidence interval in frequentist statistics was defined as posterior probability that the parameter lies within the interval:

$$\int_{-\infty}^{A} p\left(\mathbf{\Lambda} \mid Y\right) d\mathbf{\Lambda} = \int_{B}^{\infty} p\left(\mathbf{\Lambda} \mid Y\right) d\mathbf{\Lambda} = \alpha / 2,$$

where α is the significance level. Instead of analytically estimating the CI, the confidence interval for $\left(\hat{\mu}_k, \hat{\sigma}_k^2\right)$ was numerically estimated from quartiles of posterior distribution.

Simulation

To demonstrate the performance of the proposed model, two simulations were performed. Simulation I was used to evaluate model performance on complete data. It was generated from three GMMs with respective means -1, 0 and 1 and variances of 0.360, 0.640 and 0.040, respectively. A total of 150 simulated data were evenly distributed (mixing proportion was 1/3) to the three components. Simulation II was used to evaluate model performance on a truncated distribution with similar shape to Figure 2.1. A truncated Gaussian mixture with two mixture components was simulated. Zero mean was assigned to both components. The first mixture

component had mixing proportion π_1 and variance σ_1^2 of 0.800 and 0.023, respectively; the second mixture component had π_2 and σ_2^2 of 0.200 and 0.360, respectively. Truncation points were arbitrarily set to ± 0.1. In both simulations, the SE τ_i was generated from a uniform distribution [0, 0.01].

2.3.3 Results

QTL additive effects

The distributions of observed additive QTL effects from the various data sets varied with respect to the magnitude of effects and the variance of the distribution (Figure 2.1). In this study, we arbitarilly set $0.5\sigma_p$ as the threshold for QTL effects considered to be large. The positive or negative sign of the effects indicates which parent is contributing the favorable additive effect allele. In corn data I, a large number of small QTL effects were detected, with an average value of $0.2\sigma_p$ (Table 2.7a) whereas a few small QTL effects plus some large effects were detected in corn data II with the mean of $0.4\sigma_p$ (Table 2.7a). In contrast with corn, no effects around zero were detected in rice and wheat data (Figure 2.1c, d), with means of $0.45\sigma_p$ and $0.3\sigma_p$, respectively (Table 2.7a). Note that above means were calculated from the absolute QTL additive effects as only the magnitude of effects (not parent of origin) was emphasized.

DPGMM fitted normal distributions to the additive effects. The number of mixture components was inferred by the mode of posterior distribution with regard to cluster indicator c_i. The histogram of cluster sizes of four datasets clearly suggested fitting all data into one cluster (Figure 2.2). For each of the four data sets, the fitted distribution was overlaid on the histogram of "doubled" data (Figure 2.3). All distributions have zero mean; variances differed (Table 2.7b). The lowest and highest variances came from distribution of corn data I (0.044, with 95% BCI of 0.034 to 0.059) and rice data (0.220, with 95% BCI of 0.163 to 0.297), respectively.

QTL dominance coefficients

Observed dominance coefficients obtained from meta-analysis of 5 mapping studies varied from lower than -2.0 to over 2.0 (Figure 2.4), suggesting that all classes of dominance were represented among the traits measured. Around 50% of the QTLs (50 out of 101) displayed *d/a* ratio in the range of -0.5 to 0.5, indicating partial recessivity, additivity, and partial dominance gene action. Approximately 25% of the QTLs exhibited either partial dominant or dominant gene action ($0.5 < d/a < 1.25$) or partial recessive or recessive gene action ($-1.25 < d/a < -0.5$). Furthermore, 25% of the QTLs exhibited apparent overdominance (> 1.25) or underdominance (< -1.25) gene action. The extreme cases depicted expression of overdominance ($d/a > 2$) in grain yield and yield components among progeny from the cross between X178 and B73 (15 out of 16 QTLs) and expression of underdominance ($d/a < 2$) in kernel protein content among progeny derived from the cross between Ce03005 and B73.

Dominance coefficients were fitted with the normal distribution using DPGMM. Parameters for the distribution (i.e. number of clusters (K), mean (μ) and variance (σ^2)) were estimated through Gibbs sampling (Figure 2.5a). Setting 80,000 Markov Chain Monte Carlo (MCMC) iterations as burn-in period, we employed the remaining 20,000 steps as the samples of the posterior distribution. The mode of posterior distribution with regard to K is one, suggesting that all data could be fitted to a single component (Figure 2.5b shows the estimated distribution overlaid on the density plot of observed data). The estimated distribution mean was 0.152 with 95% BCI of 0.055 to 0.237, with variance 0.329 with 95% Bayesian confidence interval to be 0.193 to 0.542.

Figure 2.2 Histogram of the cluster sizes in four data sets: a) corn data I; b) corn data II; c) rice data; and d) wheat data.

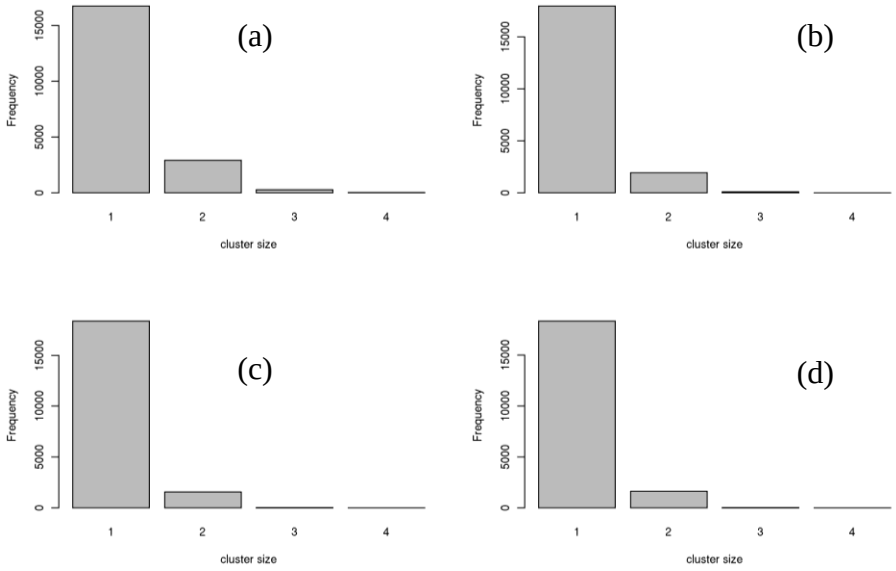


Table 2.7 a) The mean of the absolute values of the observed QTL additive effects and fitted normal distribution; b) Estimates and Bayesian confidence interval for parameters in distribution of additive effects. Values are in unit of phenotypic standard deviation.

(a)

Data sets	Observed mean	Fitted mean
Corn I	0.2	0.167
Corn II	0.4	0.306
Rice	0.45	0.374
Wheat	0.3	0.247

(b)

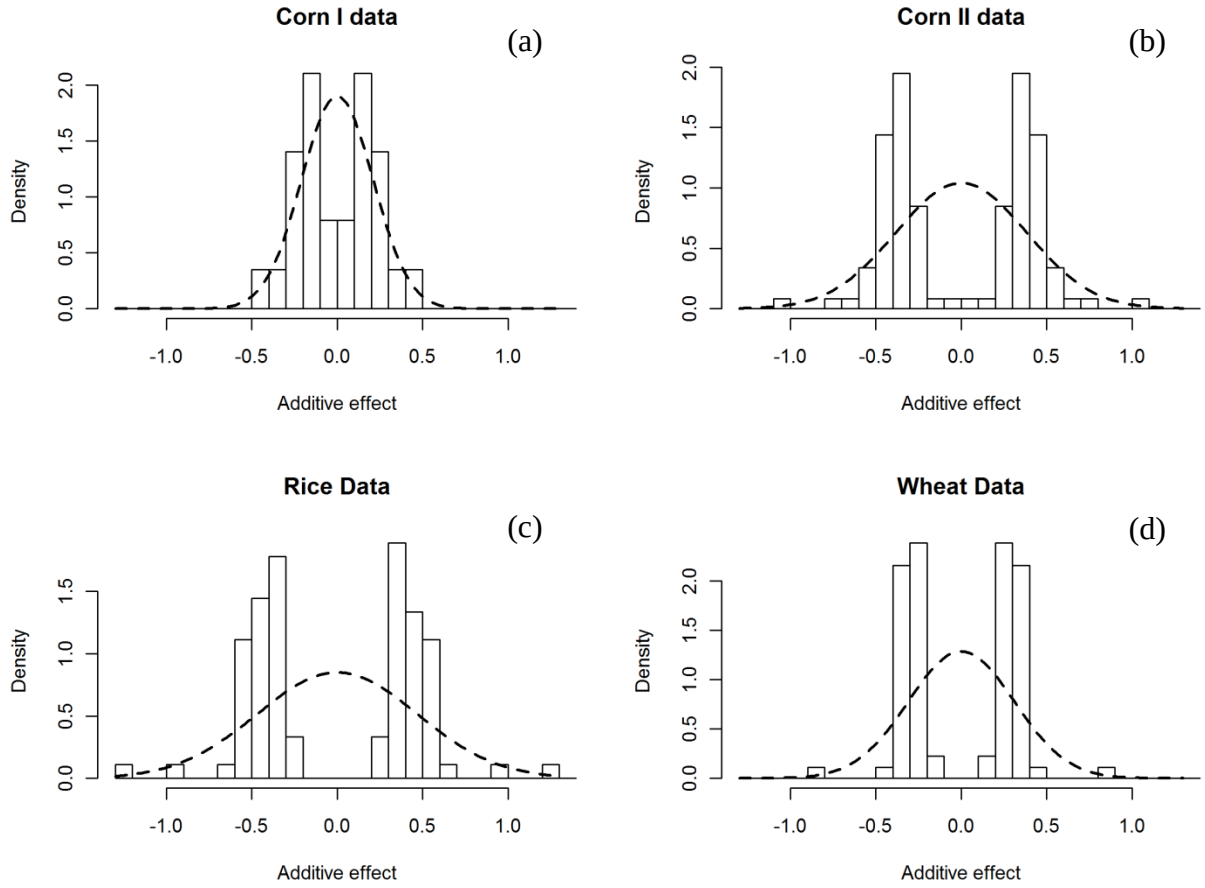
Data sets	Parameters	Posterior mean	Bayesian confidence interval	
			2.50%	97.50%
Corn I	$\hat{\sigma}_1^2$	0.044	0.034	0.059
Corn II	$\hat{\sigma}_1^2$	0.147	0.110	0.194
Rice	$\hat{\sigma}_1^2$	0.220	0.163	0.297
Wheat	$\hat{\sigma}_1^2$	0.096	0.071	0.130

Simulation

To further verify the accuracy of estimated parameters in fitting the distribution of QTL additive effects and dominance coefficients using DPGMM, two simulations were performed. In Simulation I, three components with means -1, 0, and 1, and variances of 0.360, 0.640, and 0.040, respectively, result in a histogram of genetic effects from which it is difficult to infer the number of mixture components visually (Figure 2.6a). Simulation II based on a truncated mixture normal featuring two mixture components with both with mean zero and variances of 0.023 and 0.360, respectively, produced the histogram displayed in Figure 2.6b. Data points with values from -0.1 to 0.1 had been removed to emulate the scenario in Figure 2.1.

In Simulation I, DPGMM clearly fitted the data to three clusters with estimated values close to true values (Table 2.8a). DPGMM predicted accurately the mean and variance of Clusters 1 and 3, although miss assignments of cluster membership were observed. In contrast, the mixing proportion of Cluster 2 was estimated precisely; however, certain deviations from the true mean and variance were observed. In Simulation II, all parameters were estimated accurately, except for the variance of Cluster 1, which was estimated at 0.251 versus the true value of 0.023 (Table 2.8b).

Figure 2.3 Fitted normal distribution to the QTL additive effects in a) corn data I; b) corn data II; c) rice data; and d) wheat data. Values are in unit of phenotypic standard deviation.



2.3.4 Discussion

Evidence from QTL studies in livestock has favored the leptokurtic distribution of QTL effects with many loci of small effects and few loci of large effects (Bennewitz and Meuwissen 2010; Bost et al. 2001; Bost et al. 1999; Hayes and Goddard 2001). In plants, Edwards et al. (1987) found an L-shaped distribution of QTL effects in maize for 25 traits comprising yield and domestication traits.

Based on previous results, we anticipated two mixture components to best fit additive QTL effects, one component with high mixing proportion and small variance to represent a large number of small effects and the other component having relatively large variance and low mixing proportion to represent a small number of large effects. However, only one component was fitted to all four data sets, which might be caused by two reasons. First of all, the distribution of additive QTL effects in corn comes from specific classes of traits: yield-related traits were included in corn data I and domestication traits were included in corn data II. The variance of the fitted distribution with corn data I is as small as 0.044 while the variance with corn data II is three times larger, suggesting differences in genetic architecture between these two data sets. The other reason might be from the problem of truncated data. The absence of near-zero effects in corn data II, rice data, and wheat data significantly reduced the ability to represent all the small effects in these distributions, creating a gap between the fitted distributions and the observed density plots, especially for values around zero (Figure 2.3). The above phenomenon is further reflected by the lower mean of the fitted normal distribution compared to the distribution of observed additive QTL effects in all four data sets; corn data I which retained many near-zero effects showed the least (Table 2.7a).

Table 2.8 True vs. estimated (hat) parameters in a) simulation I and b) simulation II. π_k is the mixing proportion in the k^{th} cluster, and μ_k and σ_k^2 are the mean and variance of k^{th} mixture component, respectively.

a)

	Cluster1	**Cluster2**	**Cluster3**
π_k	0.333	0.333	0.333
$\hat{\pi}_k$	0.487	0.367	0.147
μ_k	1.000	0.000	-1.000
$\hat{\mu}_k$	0.841	-0.673	-1.041
σ_k^2	0.360	0.640	0.040
$\hat{\sigma}_k^2$	0.312	0.303	0.012

b)

	Cluster1	**Cluster2**
π_k	0.800	0.200
$\hat{\pi}_k$	0.912	0.089
σ_k^2	0.023	0.360
$\hat{\sigma}_k^2$	0.251	0.382

Our results for distribution of dominance coefficients are in accordance with two previous reports in maize (Edwards et al. 1987; Stuber et al. 1987). The gene action profile found in this study has high similarity to the one found by Edwards et al. (1987). For example, both studies found roughly 50% of the investigated QTLs were additive or partially dominant, leading to a high density of dominance coefficients around zero (Figure 2.5b). This result also suggests

that the distribution of dominance coefficients more accurately reflects the true distribution because during QTL mapping, any significant effect, either additive or dominant, would result in a QTL to be retained in the model, which theoretically doubled the size of small valued data. Because of high frequency of small effects, a normal distribution with a positive mean fit the data smoothly. Kacser and Burns (1981) also concluded that dominance coefficients tend to have positive direction. Meanwhile, the majority of the overdominance effects were associated with grain yield, not surprisingly since this trait reflects heterosis that has been exploited during selection (Stuber et al. 1987). Furthermore, Edwards et al. (1987) interpreted overdominance as an over-estimation of the *d/a* ratio caused by linkage of more than one QTL to a marker locus, with each QTL expressing partial dominance.

Figure 2.4 Histogram of observed dominance coefficients from meta-analysis using five mapping studies.

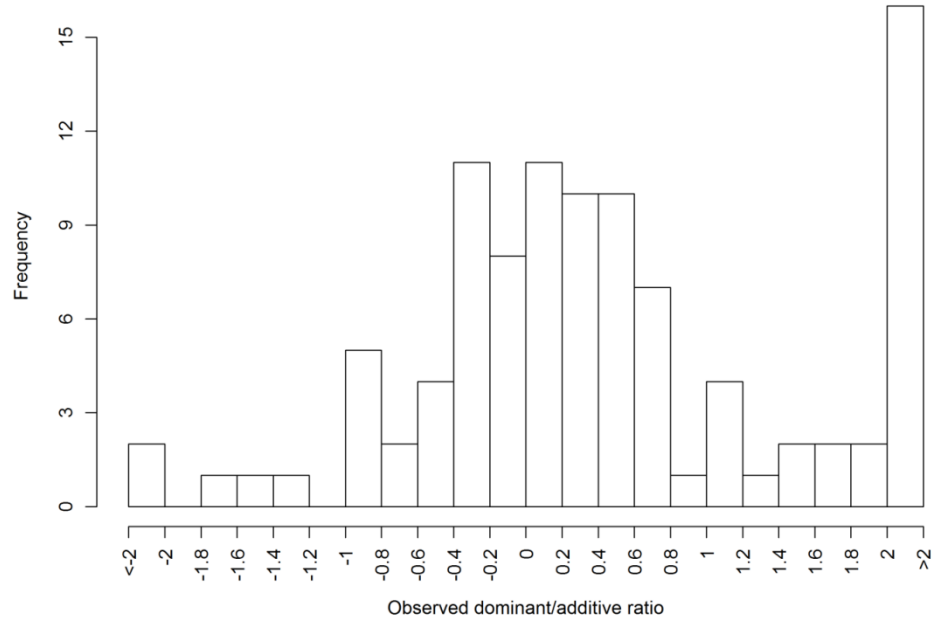
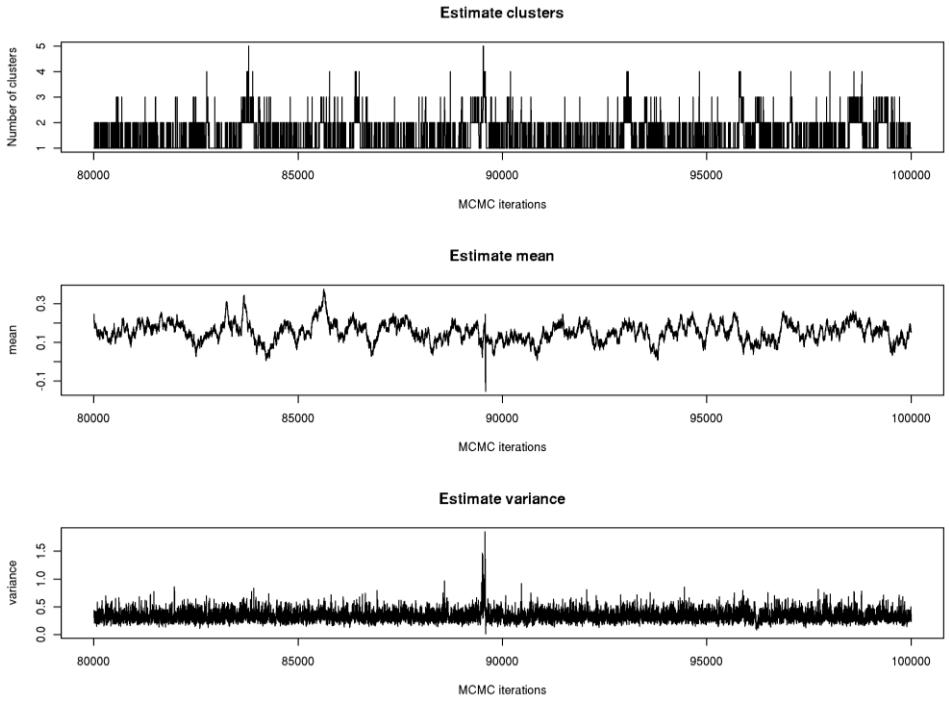
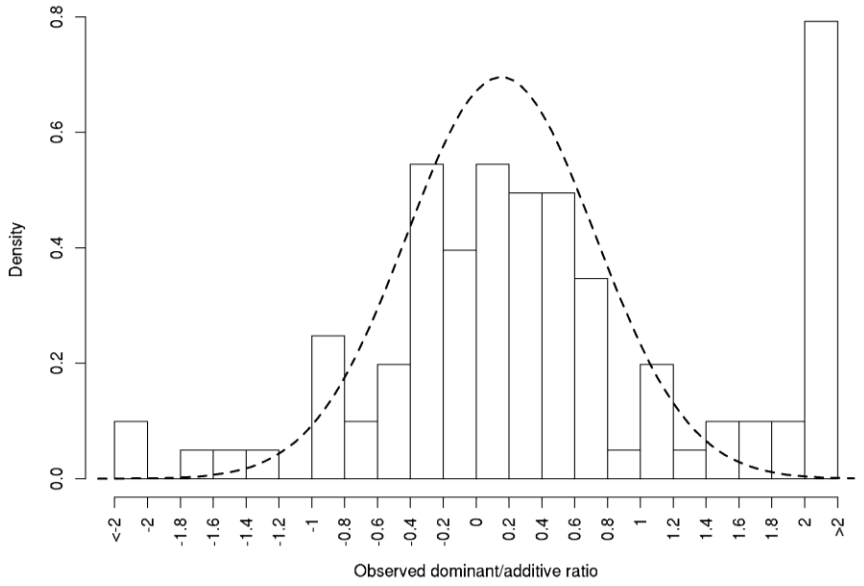


Figure 2.5 (a) Estimation of cluster size. (b) Fitted normal distribution to the dominance coefficient. The estimated mean was 0.152 with 95% Bayesian confidence interval to be 0.055 and 0.237. The estimated variance was 0.329 with 95% Bayesian confidence interval to be 0.193 and 0.542.

(a)



(b)



The distribution of dominance coefficients was derived from a meta-analysis across five QTL mapping studies. Therefore, the Beavis effect (Beavis 1998) which was defined as the tendency to overestimate QTL effects might be raised as an issue during meta-analysis. Otto and Jones (2000) suggested that given the small population size for each QTL mapping study, the number of loci was estimated downward while the QTL effects would be estimated upward. Later, Xu (2003b) investigated the statistical side of the Beavis effect and derived a formula to correct it for meta-analysis, mainly to guard against repeated reporting of the same QTL across experiments. However, the solution only works when the sample size is sufficiently large enough. Secondly, there was not a great deal of overlap among populations for traits analyzed (Table 2.6). Moreover, compared to the dominance coefficients obtained from Edwards et al. (1987) who adopted data from single population, a similar pattern of distribution was observed. Therefore, we think the Beavis effect does not hugely impact our results in this study and no correction is needed for the data.

In this study, we employed a new model, namely DPGMM, to investigate the distribution of QTL additive effects and dominance coefficients in the form of mixtures of normals. Although similar to the fitting of a mixtures of normals using a modified EM algorithm (Bennewitz and Meuwissen (2010), this approach differed primarily in the way of dealing with cluster size (K). With use of a finite mixture model and certain clustering algorithms like the EM algorithm, the number of components needs to be preset and later decided under certain circumstances, or selected by some measure, *e.g.* Akaike information criterion (AIC) and Bayesian information criterion (BIC). The optimum cluster size (K) will strike a balance between maximum data compression (assigning all data to one component) and maximum accuracy (allowing the number of clusters equal sample size). By employing the Dirichlet process, which

assumes infinite number of mixture components, model selection issue was avoided. In the present study, the Dirichlet process was represented via the Chinese Restaurant Process (CRP) (Aldous 1985). Using the CRP, a data point was assigned either to a currently occupied mixture component with probability proportional to the number of data already held in that cluster, or to a new cluster with probability proportional to the concentration parameter. By the same token, in each iteration of Gibbs sampling , the cluster indicators were also updated along with parameters like the mean and variance. As such, DPGMM fits the data distribution and explores the potential number of mixture components simultaneously.

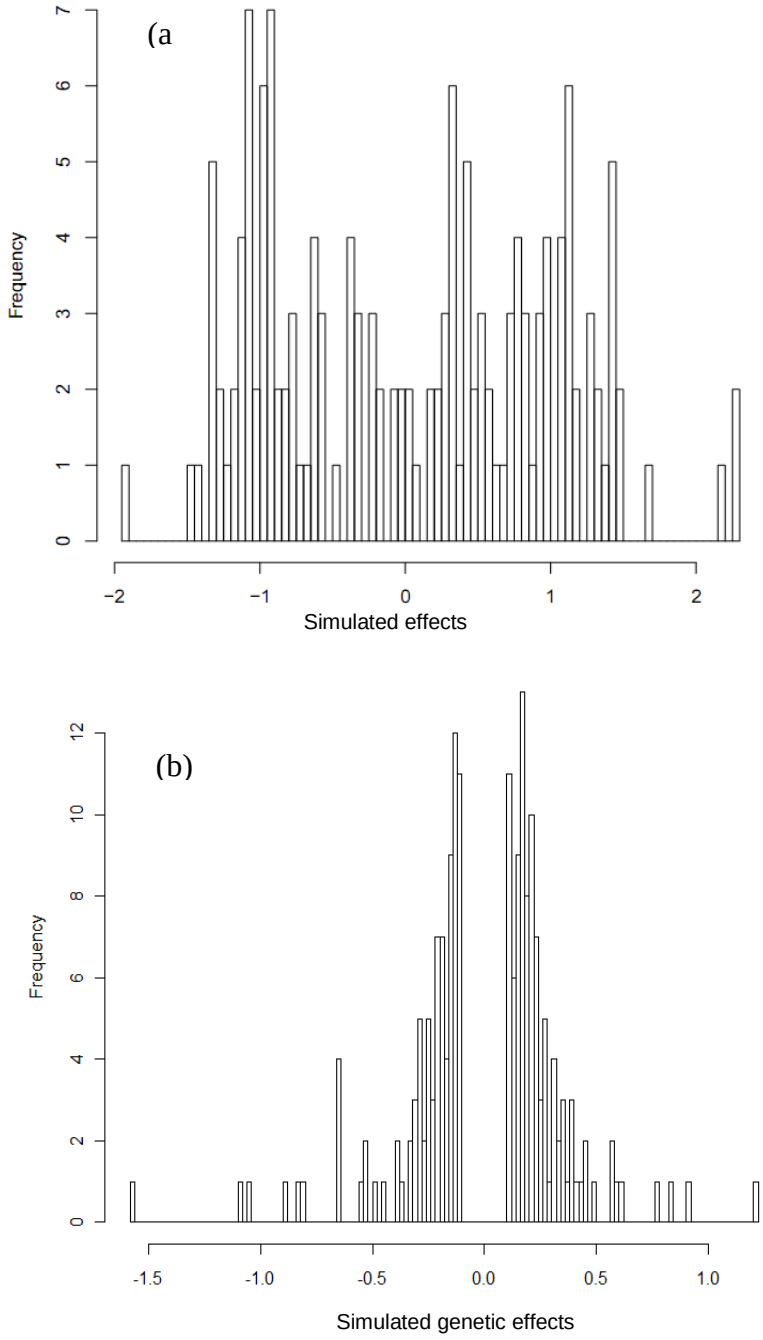
Model performance on complete and truncated data was also illustrated. Given a complete set of data, DPGMM could clearly assign membership to respective clusters with small prediction errors (Figure 2.6a, Table 2.8). In the case of truncated data, DPGMM was still effective in predicting the correct number of mixture components and estimating the variance of components with greater variability; however, DPGMM was less effective in estimating variance of components with small differences among cluster members. As shown in Table 2.8, the deviation of estimated variance (0.251) from true value (0.023) was somewhat large and might be attributed to the loss of small valued data in the sample. The above result is in accord with the conclusion that small effects would be missed easily with a mixture model (Bennewitz and Meuwissen 2010).

In short, we fitted the distribution of QTL additive effects and dominance coefficients using DPGMM which simultaneously predicted the number of mixture components and estimated the parameters. The result that the distribution of dominance coefficients was fitted to a normal distribution with a positive mean conformed to previous studies. Four separate normal distributions were derived for the QTL additive effects of four data sets. These results might be

interpreted with caution since the predicative ability of the model may have been impacted by the censored samples of identified QTL. Censorship can be reduced through use of larger population sizes, denser marker sets, more precise experimental designs in the collection of phenotypic data in the identification of QTL impacting expression of traits of interest (Beavis 1998; Xu 2003b). For example, by applying certain high resolution mapping methods, such as genome wide association mapping using high density SNP markers, a greater number near-zero sized genetic effects could be tracked (Yu et al. 2006). Furthermore, it would also be helpful to investigate the distribution of QTL effects on a trait-by-trait basis rather than group QTLs across traits as was done here.

The distributions of QTL additive and dominance effects highlighted through this study shed further light on the genetic architecture of important traits of interest. These results could be used to update the genome model used in genetic simulation for more accurate representation of QTL effects for important traits of interest (Sun et al. 2011). This would allow breeding strategies to be compared with greater precision and thus provide more value through genetic simulation; breeding decisions based on the simulation could be made with more confidence. In addition, results could aid in estimating the number of loci (or genetic factors) controlling the target trait (Zeng 1992), and serve as prior knowledge for calculation of genomic-based estimates of breeding value (Meuwissen et al. 2001).

Figure 2.6 Histograms of simulated effects from Gaussian mixtures with (a) (n=150) three components having mean of -1, 0 and 1, and variance of 0.36, 0.64 and 0.04, respectively; and (b) (n=300) two components having zero means and variance of 0.025 and 0.36, respectively. The mixing proportions were set to 1/3 for all three components in (a) and 0.8 and 0.2, respectively in (b). Distribution in (b) is truncated at points -0.1 and 0.1.



Chapter 3 - Genomics-based prediction of performance of quantitative traits involving epistasis using a nonparametric method

3.1 Introduction

The estimation of breeding values to facilitate choice of parents is a central problem in plant breeding. Fernando and Grossman (1989) first demonstrated the utility of molecular marker data to estimate breeding values in livestock species. These were data involving very few markers. Due to increasingly developed genotyping and sequencing technologies, densely spaced genome-wide SNP (single nucleotide polymorphism) data, involving tens or hundreds of thousands of markers, are now available for a number of crops. The genome-wide markers can be used as 'predictors' to achieve high accuracy in estimating breeding values. However, problems like high dimensionality and multicollinearity emerge when the number of predictors is very large and exceeds the number of records. Therefore, statistical methods to effectively address those issues are urgently needed.

Meuwissen *et al.* (2001) proposed a procedure called Genomic Selection (GS), which uses genome-wide markers to estimate breeding value. By regressing phenotypes on genome-wide markers via linear regression, this method can model high-dimensional predictors. Then, a shrinkage method can be applied to effectively 'shrink' the effect of multicollinearity and to provide stable parameter estimates (Gruber 1998). Utilizing the two techniques, this approach can generate information about genomic regions that may affect the trait of interest. Since then, several other shrinkage methods have been developed to estimate breeding values (de los Campos et al. 2009; Xu 2003a). These methods are primarily based on linear models, which is easy to interpret and able to fit to the data without overfitting. However, the relationship between breeding value and genetic markers is likely to be more complex than a simple linear

relationship, in particular when accounting for epistasis which can be an important source of genetic variation (Dudley and Johnson 2009). Furthermore, strong genetic assumptions are needed to statistically decompose epistatic variance in those linear models (Cockerham 1954), and the fact that biological basis of epistasis is not well understood makes accommodation for it in the genetic model even more difficult. To address these issues, model-free or so-called nonparametric methods which side-step linearity and require fewer genetic assumptions have gained more and more attention (Gianola et al. 2006; Gianola and van Kaam 2008; Gonzalez-Recio et al. 2008).

Gianola et al. (2006) and Gianola and van Kaam (2008) first proposed reproducing kernel Hilbert spaces (RKHS) regression for estimating breeding values with genomic data and capturing epistatic interactions. The key idea of the RKHS methods is to replace the original marker values with nonlinear transformed markers through so-called 'basis functions'. After transformation, a new space of predictors is formed and can be used in regression. In reproducing kernel Hilbert spaces, the base functions are reproducing kernels, which vary according to different inner products defined in the RKHS. Gianola and van Kaam (2008) proposed using Gaussian kernel suggested by Mallick et al. (2005) as a reproducing kernel. Besides RKHS methods, Bennewitz et al. (2009) also explored the use of a kernel method which originated from Nadaraya-Watson kernel regression (Nadaraya 1964; Watson 1964) to estimate breeding values. However, the kernel methods incur substantial bias when applied to high dimensional regression with interactions (Fan 1996).

With tens of thousands of markers in the model, the fitting of RKHS models is computationally expensive, even infeasible. The resulting algorithm is unstable and error-prone. One solution is to bring down the dimensionality of predictors through usage of a dimension

reduction method such as principal component analysis (Macciotta et al. 2010). Alternatively, one can increase prediction accuracy by filtering out 'noisy' markers. For the latter, Macciotta et al. (2009) and Schulz-Streeck et al. (2011) assigned a *p*-value to each marker through univariate linear regression and used an empirical threshold to remove markers without strong signal. And Long et al. (2007) used two steps called "filter" and "wrapper" to select SNPs. Supervised principal component analysis (SPCA) (Bair et al. 2006) offers both dimension reduction and background noise reduction and is a good choice to supplement RKHS regression.

In this study, we for the first time combine SPCA and RKHS regression to develop a two-step method (pRKHS) to estimate breeding value and predict performance. In step one, we preselect genetic markers highly correlated with the phenotype, and perform principal component analysis on the reduced marker subset. In step two, we use significant principal components as predictors in a smoothing spline ANOVA model to conduct the RKHS regression. Smoothing spline ANOVA in RKHS (Gu 2002; Wahba 1990), categorized as functional data analysis, was developed to account for nonlinear and nonadditive features in predictors. The model is fitted using a penalized least squares method, where goodness-of-fit is measured by the least squares and model complexity is dictated by a penalty. The trade-off between goodness-of-fit and model complexity is controlled by smoothing parameters, which are selected by data-driven generalized cross-validation (GCV). The pRKHS method is developed in two versions: pRKHS-NE, which accounts for only additive effects, and pRKHS-E, which includes additive-by-additive interaction effects as well as additive effects in the model. The pRKHS versions are evaluated for predictive ability in simulated genetic scenarios and confirmed in real life scenarios for utility using actual data from corn and barley. The pRKHS

versions are also compared in performance with shrinkage methods that use all markers genome-wide, specifically RR-BLUP, BayesA, and BayesB (Meuwissen et al. 2001).

3.2 Materials and methods

Simulation

The breeding scheme for maize line development outlined by (Bernardo and Yu 2007) was used in the simulation of a number of plant breeding scenarios. Specifically, two unrelated inbreds were crossed to produce an F_1 population, from which N doubled haploid (DH) lines (Cycle 0) were generated and crossed to a common tester. Testcross performance data and genotypes of Cycle 0 lines were used to train the model. Based on Cycle 0 testcross phenotypes, N_{sel} lines were selected to randomly mate for two generations to produce N Cycle 1 lines. Genotypes of Cycle 1 lines were used to estimate testcross phenotypes using fitted model. The marker data were coded as $z_{ij} = 1$, if *j*th marker locus in *i*th individual was homozygous for marker allele from parental Inbred 1, and $z_{ij} = -1$ if homozygous for marker allele from parental Inbred 2. N and N_{sel} values were set as 144 and 8, respectively, according to Bernardo and Yu (2007).

The genome model for simulation was constructed according to the published maize ISU–IBM genetic map, with a total of 1788 cM (Fu et al. 2006), with recombination computed using the Kosambi map function (Kosambi 1944). One hundred QTLs were randomly positioned across the genome. Markers were evenly spaced on the chromosome at 1cM interval. Both QTLs and markers were assumed to be bi-allelic. The genotypic value for *i*th individual was calculated according to (Cockerham 1954)

$$G_i = \sum_{k=1}^{L} a_k u_{ik} + \sum_{k \neq l} \beta_{kl} u_{il} u_{ik} \ .$$

Design element u_k is defined according to the general two-allele model (G2A, (Zeng et al. 2005) as

$$u_k = \begin{cases} 2(1-p) & k\text{th QTL genotype is QQ} \\ 1-2p & k\text{th QTL genotype is Qq} \\ -2p & k\text{th QTL genotype is qq} \end{cases},$$

where p is the allele frequency of Q. Parameter a_k is kth QTL's additive effect and β_{kl} is the epistatic interaction effect between kth and lth QTL. In this case, β_{kl} indicates additive by additive interaction. Furthermore, a_k was sampled from geometric series $\left(\frac{L-1}{L+1}\right)^k$ (Bernardo and Yu 2007; Lande and Thompson 1990), where L equals the total number of QTL positioned throughout the genome. The direction of effect was randomly assigned to each QTL, leading to random coupling and repulsion linkages. Epistatic effect β_{kl} was sampled from gamma distribution

$$\beta_{kl} \sim Gamma(0.2,10)\delta_{x\le P} \text{ and } \delta = \begin{cases} 1 & x \le P \\ 0 & x > P \end{cases},$$

where δ is the indicator function. The extent of epistasis was specified by assigning the proportion (P) of total epistatic interactions with nonzero effect. Three levels of epistasis were considered: P = 0, 0.1 and 0.5, representing no, low, and high epistasis, respectively. The G2A model (Zeng et al. 2005) was chosen to model QTL due to its orthogonal property, which links genetic variance partition directly to genetic effect partition. The genetic variance was therefore calculated from the sample variance of genotypic values. Random nongenetic effects were added to the genotypic values to generate phenotypic values in proportion to the heritability (*i.e.* four heritability levels were considered: 0.1, 0.2, 0.4 and 0.8).

Real data

To evaluate the predictive ability under real life scenarios, data reported by Crossa *et al.* (2010) on 284 maize lines genotyped with 1148 SNPs and phenotyped for anthesis-silking interval (ASI) were utilized. In addition, two barley datasets generated from North Dakota State two-rowed (N2) breeding program, with trial name of Expt41_2007_Langdon and Expt41_2008_Langdon, respectively, from The Hordeum Toolbox (http://wheat.pw.usda.gov/tht/) were utilized. Only entries with phenotypic observations for both grain yield and plant height from the same location were used to avoid confounding genotype with environment. Each trial contained a different set of 96 lines for a total of 192 unique lines across the two years. There were 2161 SNPs for the 2007 dataset and 2029 SNPs for the 2008 dataset, among which 1875 markers were shared between two years.

After filtering out markers with minor allele frequency (*i.e.* smaller than 0.05), 1148 and 1642 SNPs were retained for maize and barley data, respectively. SNPs were bi-allelic and the dummy variable for marker data is defined as $z_{ij} = 1$ for A_1A_1, $z_{ij} = 0$ for A_1A_2 and $z_{ij} = -1$ for A_2A_2. For SNP data from BarleyCAP, genotypes '1:1', '2:2', and '1:2' were considered as A_1A_1, A_2A_2, and A_1A_2, respectively. Although the type of marker data is discrete, it is treated as continuous vector of covariates. Missing markers were imputed by averaging marker scores across all lines of that marker. Missing phenotypes were imputed using k-nearest-neighbor (KNN) algorithm.

Statistical methods

Features from SPCA and RKHS regression were combined to develop the new method, pRKHS. First, SPCA was applied to reduce the high dimensionality represented by the markers and to decrease 'noise'. Steps to apply SPCA included:

a) Computing the regression coefficient for each marker on a single marker basis,

b) Ranking markers by the absolute value of their regression coefficients and selecting a defined number of the top ranked markers to form a marker subset (MS) with which to construct the reduced data matrix,

c) Performing principal component analysis using the reduced data matrix to generate resulting PCs, referred to as supervised principal components (SPCs).

SPCs explaining 70% of the data matrix variance were then selected as independent variables to fit a smoothing spline ANOVA model in reproducing kernel Hilbert spaces (Gu 2002). Two versions of the new method (pRKHS-NE and pRKHS-E) were proposed to account for various levels of epistasis.

1) pRKHS-NE: All selected SPCs were included in the model as main effects. No interactions were included.

2) pRKHS-E: Main effects and two-way, additive-by-additive interactions were included in the model, specifying the level of epistasis. For example, when epistasis was specified as 0.1, then 10% of the epistasis interaction effects were considered to be nonzero. To prevent high dimensionality, each variable and their pair-wise interactions were tested for significance and selected using non-parametric model diagnostics tools, *i.e.* cosine value (CoV) (Gu 2002), which corresponds to F-statistics in a parametric regression model *i.e*. a high CoV corresponds to a high level of statistical significance. Main effects with CoV larger than 0.05 were retained in the model. To determine the tolerance for various different levels of epistatic interactions, a series of CoV *i.e.* 0.3, 0.25 and 0.2 were considered.

We modeled the phenotype and SPCs using a general nonparametric model on domain $\mathcal{X}$ = [0, 1], which can be written as

$$Y_i = \eta(\mathbf{x}_i) + \varepsilon_i,$$

where Y_i is the phenotype of ith individual, $\mathbf{x}_i = \left(x_i^{(1)}, x_i^{(2)}..., x_i^{(K)}\right) \in \mathcal{X}$ is the vector of k SPCs $x_i^{(1)}, x_i^{(2)}..., x_i^{(K)}$ of ith individual, and $\varepsilon_i \sim N\left(0, \sigma^2\right)$ is error term for ith individual. Analogous to classical ANOVA in linear models, a functional ANOVA decomposition could be written as,

$$\eta(\mathbf{x}) = \eta_0 + \sum_{j=1}^{K} \eta_j\left(x^{(j)}\right) + \sum_{j=1}^{K}\sum_{m=j+1}^{K} \eta_{jm}\left(x^{(j)}, x^{(m)}\right) + \text{all higher order of interactions},$$

where η_0 is constant, η_j's are the main effects, and η_{jm}'s are the two-way interactions with $\left(x^{(j)}, x^{(m)}\right)$ on product domain $\mathcal{X} \times \mathcal{X} = [0, 1]^2$. Each η_j was estimated in a RKHS $\mathcal{H}_j$, and each η_{jm} was estimated in the tensor product RKHS $\mathcal{H}_{jm} = \mathcal{H}_j \otimes \mathcal{H}_m$, and so on (Gu 2002). Let $(f, g)_j$ be the inner product in $\mathcal{H}_j$, $(f, g)_{jm}$ be the inner product in $\mathcal{H}_{jm}$, and so on. $\eta(\mathbf{x})$ was thus estimated by a penalized least squares in the RKHS

$\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus ... \oplus \mathcal{H}_K \oplus \mathcal{H}_{12} \oplus ... \oplus \mathcal{H}_{(K-1)K} \oplus ...,$

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \eta(\mathbf{x}_i)\right)^2 + \lambda\mathcal{J}\left(\eta\right), \qquad (1)$$

where $\mathcal{J}\left(f\right) = \theta_1^{-1}(f, f)_1 + ... + \theta_K^{-1}(f, f)_K + \theta_{12}^{-1}(f, f)_{12} + ... + \theta_{(K-1)K}^{-1}(f, f)_{(K-1)K} + ...$, and θ_βs are the inter-space inner product rescaling parameters, and λ is a smoothing parameter. The first term $\sum_{i=1}^{n}\left(Y_i - \eta(\mathbf{x}_i)\right)^2$ measures the goodness-of-fit, the second term $\lambda\mathcal{J}\left(\eta\right)$ penalizes for smoothness of the η, and the smoothing parameter λ strikes the balance between the goodness-of-fit and smoothness of the η. The subspaces $\mathcal{H}_\beta s$ form two large subspaces, 1) $\mathcal{N}_\mathcal{J} = \left\{\eta : \mathcal{J}\left(n\right) = 0\right\}$, which is the null space of $\mathcal{J}\left(\eta\right)$, and 2) $\mathcal{H} \ominus \mathcal{N}_\mathcal{J}$, with the reproducing kernel $R_\mathcal{J}(\cdot, \cdot)$.

The minimizer of (1) has expression

$$\eta(x) = \sum_{v=1}^{m} d_v \phi_v(x) + \sum_{j=1}^{n} c_j R_{\mathcal{J}}(x_j, x), \tag{2}$$

where $\{\phi_1, \ldots, \phi_m\}$ is the basis of null space $\mathcal{N}_{\mathcal{J}}$ and coefficients d_v and c_j need to be estimated from data.

In the pRKHS-NE version, we consider $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus ... \oplus \mathcal{H}_K$, and in the pRKHS-E version, we consider $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus ... \oplus \mathcal{H}_K \oplus \mathcal{H}_{12} \oplus ... \oplus \mathcal{H}_{(K-1)K}$. In the computation, any continuous SPC $x^{(j)}$ is scaled onto [0, 1] and choose $\mathcal{H}_j$ to be $\mathcal{H}_j = \left\{ \eta_j : \int_0^1 \left[\eta_j''(x^{(j)}) \right]^2 dx < \infty \right\}$.

When endowed with a certain inner product, $\mathcal{H}_j$ has the reproducing kernel (RK):

$$1 + R_0(s,t) + R_1(s,t),$$

where $R_0(s,t) = k_1(s)k_1(t)$ and $R_1(s,t) = k_2(s)k_2(t) - k_4(|s-t|)$, with $k_1(t) = t - 0.5$, $k_2(t) = \frac{1}{2}\left(k_1^2(t) - \frac{1}{12}\right)$, and $k_4(t) = \frac{1}{24}\left(k_1^4(t) - \frac{k_1^2(t)}{2} + \frac{7}{240}\right)$. The RK for $\mathcal{H}_{jm}$ is

$$\left(1 + R_0(s,t) + R_1(s,t)\right)\left(1 + R_0(s,t) + R_1(s,t)\right).$$

A refined leave-one-out cross-validation procedure called generalized cross-validation (GCV) was used to choose values for λ and $\theta_\beta s$ (Gu 2002).

pRKHS-E and pRKHS-NE were compared to three shrinkage methods: RR-BLUP, BayesA and BayesB (Meuwissen et al. 2001). The general model was written as $\mathbf{y} = \mathbf{l}\mu + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{l}$ is a vector filled with ones, $\mathbf{X}$ is marker data matrix, μ is the fixed grand mean, $\boldsymbol{\beta}$ is the vector of marker effects, and $\boldsymbol{\varepsilon}$ is the vector of random residuals. A Gaussian prior was assigned to $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$, with $\boldsymbol{\beta} \sim N\left(0, \mathbf{I}\sigma_\beta^2\right)$ and $\boldsymbol{\varepsilon} \sim N\left(0, \mathbf{I}\sigma_e^2\right)$. RR-BLUP was implemented in a Bayesian

frame and assigns common variance to all marker effects, whereas BayesA and BayesB assigns different variances to different markers. BayesB was modified in this study to include π, the proportion of markers having no genetic variances, as another parameter in the model and assign it a uniform [0,1] prior instead of arbitrary setting (Habier et al. 2011). Variances σ_e^2 and σ_β^2 were assigned a scaled inverse chi-square distribution with scale S^2 and degree of freedom v.

Data analysis

The smoothing spline ANOVA model in pRKHS-E and pRKHS-NE was fitted using the *ssanova* function in "gss" package available in R (R Development Core Team 2011). Smoothness parameter λ, which balances the goodness-of-fit of data and smoothness of the curve, was selected by GCV. RR-BLUP, BayesA and BayesB were coded using C++, among which S_β^2 = 4.23, v_β = 0.05 and S_e^2 =1, v_e = 1. Gibbs sampler was implemented with 3 chains and 10,000 iterations for each chain to update conditional posterior distributions. The first 1,000 samples of each chain were discarded as burn-in and later thinned by 10. Convergence was checked by inspection of trace plots and Gelman-Rubin plots of error variance using "coda" package in R (Plummer et al. 2006). Samples from three chains were combined to estimate posterior means. All analyses were run on an Ubuntu Server with 2.8 GHz CPU and 16GB memory.

In simulation scenarios, ten-fold CV in Cycle 0 (C0) was used to determine the best MS and CoV for pRKHS-E(NE). Pearson correlation coefficients between estimated breeding value (EBV) and true breeding value (TBV) ($r_{EBV:TBV}$), and between EBV and phenotype (PHE) ($r_{EBV:PHE}$) were calculated and averaged across five replicated simulations.

Since TBV will never be observed in real cases, the criterion to select MS and CoV was based on $r_{EBV:PHE}$ rather than $r_{EBV:TBV}$. Given the highest $r_{EBV:PHE}$ in C0, optimum MS and CoV were

determined and used to estimate breeding values and phenotypes in Cycle 1 (C1). Across a series of MS (*i.e.* from 500 to all markers), percent of variation and number of influential markers whose loadings > 0.8*the maximum loading were extracted from top three SPCs, and number of SPC interactions with CoV being 0.2, 0.25 and 0.3 were also recorded.

In real data applications, predicted maize ASI values were based on five-fold CV since only one set of data was available. With barley, $r_{EBV:PHE}$ was computed to measure predictive ability. Expt41_2007_Langdon barley data were used for ten-fold CV to select the optimum MS and CoV for pRKHS-E(NE), which were further used to predict phenotypes of trial Expt41_2008_Langdon. Since the population size in barley was small, *i.e*. 96 per year, model fitting was repeated five times and results were reported as a mean of five. The R code for computing pRKHS EBVs and correlations with phenotype using the two barley datasets generated by North Dakota State barley breeding program (Expt41_2007_Langdon and Expt41_2008_Langdon accessed through The Hordeum Toolbox http://wheat.pw.usda.gov/tht/) is provided in the supplemental materials (LINK).

3.3 Results

Simulation results

Twelve different scenarios were considered in this study to facilitate comparison of methods given various levels of heritability and epistasis (Table 3.1, 3.2, and 3.3). Pearson correlation coefficients for EBV: PHE ($r_{EBV:PHE}$) and for EBV: TBV ($r_{EBV:TBV}$) were calculated and displayed in Tables 3.1, 3.2 and 3.3. Both ten-fold CV in C0 and prediction in C1 were used to assess the predictability of the statistical methods.

For scenarios with no epistasis, BayesB generally outperformed other methods in predictive ability (Table 3.1). BayesB provided the highest correlation between EBV and TBV in

most cases, except in C0 case at h^2=0.1, where pRKHS-NE outperformed BayesB. The values of $r_{EBV:TBV}$ for pRKHS-NE were always higher than those for pRKHS-E across both C0 and C1, in keeping with the scenario of no epistasis. The pRKHS method provided higher correlations between EBV and PHE than BayesB in four out of eight cases. Among the four cases, pRKHS-E provided the highest correlation between EBV and PHE when heritability was low to moderate (h^2 = 0.1, 0.2, and 0.4) in C0 cases, whereas pRKHS-NE outperformed BayesB only when h^2 = 0.2 in C1 cases. In no instances, did RR-BLUP or BayesA provide the highest correlations for TBV or PHE.

For scenarios with epistasis at a low level, the pRKHS method outperformed other methods in predictive ability; particularly in predicting PHE, the pRKHS method provided the highest correlation in all eight cases of C0 and C1 (Table 3.2). The values of pRKHS-E for $r_{EBV:TBV}$ were generally higher than those for RKHS-NE, but only marginally. The pRKHS method provided highest values for $r_{EBV:TBV}$ in five out of eight cases of C0 and C1, with BayesB providing the highest values in the other three cases. For the correlation with PHE, pRKHS-E exceeded pRKHS-NE in three out of four C0 cases (h^2 = 0.1, 0.2, and 0.4) and performed equally to pRKHS-NE in C1 cases. In no instances did RR-BLUP or BayesA provide the highest correlations for either TBV or PHE.

For scenarios with high epistasis, the pRKHS method, particularly pRKHS-E, outperformed other methods in predictive ability (Table 3.3). pRKHS-E provided the highest correlations for both TBV and PHE in all cases of C0 and C1 across all heritabilities. Among four C0 cases, the magnitude of the values of $r_{EBV:TBV}$ of BayesB had decreased the most from the corresponding value in low epistasis scenario at low heritabilities (h^2 =0.1 and 0.2) and pRKHS-NE and BayesA experienced the most loss of accuracy at heritability of 0.4 and 0.8, respectively.

Compared to the low epistasis scenarios (Table 3.2), values of $r_{EBV:TBV}$ for BayesB among the four C0 cases decreased substantially, suggesting a reduction of unbiasedness due to the increasing extent of epistasis. This decrease was greatest for h^2 of 0.1 and 0.2. pRKHS-NE and BayesA also experienced substantial loss of accuracy at heritability of 0.4 and 0.8, respectively. pRKHS-E showed the least amount of loss in unbiasedness based on change of $r_{EBV:TBV}$ in all four C0 cases.

The advantage of marker-based selection (MBS) over phenotypic selection (PS) can be quantified by comparing $r_{EBV:TBV}$ in C1 cases to accuracy of PS, defined as the correlation between mid-parent and offspring and measured by taking square root of half of the narrow heritability (Falconer and Mackay 1996). For heritabilities of 0.1, 0.2, 0.4 and 0.8, the accuracy of PS was estimated as 0.224, 0.316, 0.447, and 0.632, respectively. For the scenarios with no and low epistasis (Table 3.1, 3.2), all methods outperformed PS at all four heritabilities, particularly at low heritability (h^2 = 0.1 and 0.2). For the scenarios with high epistasis (Table 3.3), not all methods outperformed PS across the four heritability levels. pRKHS-E had higher performance than PS in three out of four cases (h^2 = 0.1, 0.2, and 0.8), whereas pRKHS-NE, BayesB and RR-BLUP outperformed PS only at low heritabilities (h^2 = 0.1 and 0.2).

Table 3.4 exhibits the range of values for the percentage of variation explained by the first three SPCs, the number of markers included in each of these SPCs, and the number of SPC interactions observed when the cosine threshold for selecting SPC interactions was >0.2, >0.25, and >0.3, respectively, given the series of marker subsets (from as low as 500 markers to all, *i.e.* 1798 markers) used across 12 simulation scenarios. With low marker density, *i.e.* 500 markers was used, the first, second and third SPCs explained up to 18.7%, 11.8% and 9.9%, respectively, of the marker variations. In case of utilizing full marker information, the top three SPCs only

explained as low as 9.2%, 5.5% and 5.1% of the variations (Table 3.4). Averaging across all scenarios, the first three SPC accounted for 25.4% of the marker variation and 18 SPCs were needed to explain 70% of the marker variation (Figure 3.1). The number of influential markers included in the first three SPCs, namely M_{P1}, M_{P2} and M_{P3}, varied according to the number of markers used. Using all 1798 markers in simulations, the 1st, 2nd and 3rd SPCs included a maximum of 137, 111, and 103 influential markers, respectively; using the smallest subset of markers i.e. 500 markers, the 1st, 2nd, and 3rd SPCs included as few as 61, 52, and 37 influential markers, respectively (Table 3.4). The number of SPC interactions was determined by the MS and CoV. High MS or low CoV generate large number of SPC interactions.

Table 3.1 For scenarios with no epistasis, Pearson correlation coefficients between estimated breeding value and true breeding value ($r_{EBV:TBV}$) or phenotype ($r_{EBV:PHE}$) obtained through ten-fold cross-validation with Cycle 0 (C0) and prediction of Cycle 1(C1), implemented for simulated traits with heritability of 0.1, 0.2, 0.4, 0.8, via the various statistical methods. Average correlations ± SE were obtained from five replicated simulations.

Heritability	C0 / C1	Methods	$r_{EBV:TBV}$ ± SE	$r_{EBV:PHE}$ ± SE
$h^2 = 0.1$	C0	RR-BLUP	0.490 ± 0.049	0.176 ± 0.066
	C0	Bayes-A	0.467 ± 0.053	0.172 ± 0.093
	C0	Bayes-B	0.491 ± 0.048	0.182 ± 0.075
	C0	pRKHS-E	0.436 ± 0.053	0.233 ± 0.055
	C0	pRKHS-NE	0.494 ± 0.034	0.216 ± 0.106
	C1	RR-BLUP	0.528 ± 0.036	0.199 ± 0.030
	C1	Bayes-A	0.506 ± 0.055	0.196 ± 0.023
	C1	Bayes-B	0.537 ± 0.037	0.206 ± 0.031
	C1	pRKHS-E	0.417 ± 0.119	0.157 ± 0.050
	C1	pRKHS-NE	0.497 ± 0.074	0.191 ± 0.009
$h^2 = 0.2$	C0	RR-BLUP	0.492 ± 0.126	0.260 ± 0.072
	C0	Bayes-A	0.488 ± 0.102	0.255 ± 0.047
	C0	Bayes-B	0.502 ± 0.120	0.269 ± 0.084
	C0	pRKHS-E	0.392 ± 0.137	0.349 ± 0.088
	C0	pRKHS-NE	0.497 ± 0.181	0.325 ± 0.160
	C1	RR-BLUP	0.515 ± 0.107	0.293 ± 0.055
	C1	Bayes-A	0.476 ± 0.053	0.269 ± 0.066
	C1	Bayes-B	0.524 ± 0.112	0.302 ± 0.061
	C1	pRKHS-E	0.405 ± 0.104	0.236 ± 0.018
	C1	pRKHS-NE	0.514 ± 0.078	0.313 ± 0.039
$h^2 = 0.4$	C0	RR-BLUP	0.785 ± 0.027	0.398 ± 0.034
	C0	Bayes-A	0.697 ± 0.034	0.331 ± 0.026
	C0	Bayes-B	0.799 ± 0.032	0.404 ± 0.019
	C0	pRKHS-E	0.756 ± 0.053	0.430 ± 0.017
	C0	pRKHS-NE	0.789 ± 0.039	0.423 ± 0.039
	C1	RR-BLUP	0.688 ± 0.027	0.521 ± 0.026
	C1	Bayes-A	0.603 ± 0.030	0.457 ± 0.052
	C1	Bayes-B	0.696 ± 0.031	0.529 ± 0.020
	C1	pRKHS-E	0.628 ± 0.020	0.475 ± 0.033
	C1	pRKHS-NE	0.689 ± 0.035	0.496 ± 0.029
$h^2 = 0.8$	C0	RR-BLUP	0.823 ± 0.025	0.673 ± 0.092
	C0	Bayes-A	0.759 ± 0.046	0.617 ± 0.075
	C0	Bayes-B	0.827 ± 0.029	0.679 ± 0.101
	C0	pRKHS-E	0.754 ± 0.031	0.654 ± 0.078
	C0	pRKHS-NE	0.821 ± 0.017	0.643 ± 0.078
	C1	RR-BLUP	0.747 ± 0.060	0.704 ± 0.086
	C1	Bayes-A	0.667 ± 0.119	0.631 ± 0.134
	C1	Bayes-B	0.755 ± 0.055	0.712 ± 0.084
	C1	pRKHS-E	0.652 ± 0.057	0.602 ± 0.059
	C1	pRKHS-NE	0.751 ± 0.067	0.704 ± 0.048

Table 3.2 For scenarios with a low level of epistasis (10% of the epistasis interaction effects are nonzero), Pearson correlation coefficients between estimated breeding value and true breeding value ($r_{EBV:TBV}$) or phenotype ($r_{EBV:PHE}$) obtained through ten-fold cross-validation with Cycle 0 (C0) and prediction of Cycle 1 (C1), implemented for simulated traits with heritability of 0.1, 0.2, 0.4, 0.8, via the various statistical methods. Average correlations ± SE were obtained from five replicated simulations.

Heritability	C0 / C1	Methods	$r_{EBV:TBV}$ ± SE	$r_{EBV:PHE}$ ± SE
$h^2 = 0.1$	C0	RR-BLUP	0.398 ± 0.170	0.210 ± 0.096
	C0	Bayes-A	0.391 ± 0.156	0.210 ± 0.105
	C0	Bayes-B	0.399 ± 0.186	0.190 ± 0.100
	C0	pRKHS-E	0.398 ± 0.111	0.244 ± 0.112
	C0	pRKHS-NE	0.384 ± 0.126	0.197 ± 0.094
	C1	RR-BLUP	0.298 ± 0.176	0.119 ± 0.061
	C1	Bayes-A	0.287 ± 0.158	0.111 ± 0.057
	C1	Bayes-B	0.332 ± 0.121	0.125 ± 0.076
	C1	pRKHS-E	0.342 ± 0.194	0.129 ± 0.079
	C1	pRKHS-NE	0.306 ± 0.174	0.120 ± 0.086
$h^2 = 0.2$	C0	RR-BLUP	0.465 ± 0.004	0.095 ± 0.096
	C0	Bayes-A	0.481 ± 0.063	0.101 ± 0.071
	C0	Bayes-B	0.494 ± 0.031	0.106 ± 0.099
	C0	pRKHS-E	0.475 ± 0.017	0.213 ± 0.055
	C0	pRKHS-NE	0.473 ± 0.084	0.161 ± 0.079
	C1	RR-BLUP	0.492 ± 0.081	0.295 ± 0.005
	C1	Bayes-A	0.431 ± 0.103	0.257 ± 0.041
	C1	Bayes-B	0.505 ± 0.127	0.303 ± 0.030
	C1	pRKHS-E	0.432 ± 0.012	0.299 ± 0.019
	C1	pRKHS-NE	0.493 ± 0.163	0.319 ± 0.050
$h^2 = 0.4$	C0	RR-BLUP	0.632 ± 0.068	0.421 ± 0.097
	C0	Bayes-A	0.576 ± 0.080	0.398 ± 0.069
	C0	Bayes-B	0.641 ± 0.062	0.415 ± 0.100
	C0	pRKHS-E	0.628 ± 0.060	0.435 ± 0.079
	C0	pRKHS-NE	0.643 ± 0.071	0.431 ± 0.083
	C1	RR-BLUP	0.589 ± 0.122	0.430 ± 0.103
	C1	Bayes-A	0.518 ± 0.068	0.380 ± 0.070
	C1	Bayes-B	0.583 ± 0.142	0.425 ± 0.116
	C1	pRKHS-E	0.595 ± 0.120	0.439 ± 0.112
	C1	pRKHS-NE	0.590 ± 0.095	0.431 ± 0.085
$h^2 = 0.8$	C0	RR-BLUP	0.798 ± 0.054	0.709 ± 0.072
	C0	Bayes-A	0.730 ± 0.066	0.650 ± 0.095
	C0	Bayes-B	0.812 ± 0.067	0.716 ± 0.093
	C0	pRKHS-E	0.807 ± 0.097	0.718 ± 0.083
	C0	pRKHS-NE	0.819 ± 0.066	0.721 ± 0.067
	C1	RR-BLUP	0.736 ± 0.021	0.690 ± 0.017
	C1	Bayes-A	0.676 ± 0.015	0.634 ± 0.008
	C1	Bayes-B	0.753 ± 0.021	0.705 ± 0.011
	C1	pRKHS-E	0.770 ± 0.083	0.699 ± 0.082
	C1	pRKHS-NE	0.749 ± 0.086	0.708 ± 0.097

Table 3.3 For scenarios with a low level of epistasis (50% of the epistasis interaction effects are nonzero), Pearson correlation coefficients between estimated breeding value and true breeding value ($r_{EBV:TBV}$) or phenotype ($r_{EBV:PHE}$) obtained through ten-fold cross-validation with Cycle 0 (C0) and prediction of Cycle 1 (C1), implemented for simulated traits with heritability of 0.1, 0.2, 0.4, 0.8, via the various statistical methods. Average correlations ± SE were obtained from five replicates.

Heritability	C0 / C1	Methods	$r_{EBV:TBV}$ ± SE	$r_{EBV:PHE}$ ± SE
$h^2 = 0.1$	C0	RR-BLUP	0.192 ± 0.146	0.173 ± 0.147
	C0	Bayes-A	0.183 ± 0.138	0.156 ± 0.154
	C0	Bayes-B	0.156 ± 0.117	0.139 ± 0.106
	C0	pRKHS-E	0.211 ± 0.119	0.213 ± 0.089
	C0	pRKHS-NE	0.192 ± 0.104	0.181 ± 0.096
	C1	RR-BLUP	0.326 ± 0.100	0.119 ± 0.034
	C1	Bayes-A	0.344 ± 0.080	0.129 ± 0.024
	C1	Bayes-B	0.315 ± 0.076	0.125 ± 0.031
	C1	pRKHS-E	0.349 ± 0.101	0.133 ± 0.055
	C1	pRKHS-NE	0.319 ± 0.081	0.115 ± 0.050
$h^2 = 0.2$	C0	RR-BLUP	0.234 ± 0.113	0.165 ± 0.208
	C0	Bayes-A	0.226 ± 0.113	0.168 ± 0.190
	C0	Bayes-B	0.234 ± 0.124	0.177 ± 0.237
	C0	pRKHS-E	0.320 ± 0.071	0.224 ± 0.137
	C0	pRKHS-NE	0.314 ± 0.037	0.200 ± 0.127
	C1	RR-BLUP	0.352 ± 0.116	0.184 ± 0.160
	C1	Bayes-A	0.309 ± 0.198	0.158 ± 0.150
	C1	Bayes-B	0.372 ± 0.140	0.201 ± 0.158
	C1	pRKHS-E	0.410 ± 0.105	0.206 ± 0.089
	C1	pRKHS-NE	0.330 ± 0.077	0.189 ± 0.075
$h^2 = 0.4$	C0	RR-BLUP	0.437 ± 0.067	0.262 ± 0.041
	C0	Bayes-A	0.431 ± 0.065	0.260 ± 0.074
	C0	Bayes-B	0.468 ± 0.078	0.299 ± 0.066
	C0	pRKHS-E	0.491 ± 0.075	0.326 ± 0.119
	C0	pRKHS-NE	0.439 ± 0.085	0.312 ± 0.068
	C1	RR-BLUP	0.431 ± 0.136	0.310 ± 0.071
	C1	Bayes-A	0.389 ± 0.102	0.284 ± 0.048
	C1	Bayes-B	0.438 ± 0.128	0.314 ± 0.066
	C1	pRKHS-E	0.440 ± 0.119	0.317 ± 0.052
	C1	pRKHS-NE	0.378 ± 0.108	0.271 ± 0.047
$h^2 = 0.8$	C0	RR-BLUP	0.424 ± 0.154	0.404 ± 0.138
	C0	Bayes-A	0.343 ± 0.147	0.321 ± 0.147
	C0	Bayes-B	0.441 ± 0.162	0.417 ± 0.144
	C0	pRKHS-E	0.622 ± 0.124	0.552 ± 0.135
	C0	pRKHS-NE	0.576 ± 0.177	0.497 ± 0.188
	C1	RR-BLUP	0.365 ± 0.090	0.339 ± 0.074
	C1	Bayes-A	0.349 ± 0.105	0.324 ± 0.088
	C1	Bayes-B	0.390 ± 0.077	0.361 ± 0.065
	C1	pRKHS-E	0.633 ± 0.010	0.568 ± 0.088
	C1	pRKHS-NE	0.613 ± 0.051	0.431 ± 0.028

Experimental data

In addition to the simulations, the predictive ability of each method was compared using maize and barley data reported by CIMMYT (Crossa et al. 2010) and BarleyCAP, respectively. With the maize dataset, five-fold CV was implemented with the trait anthesis-silking interval (ASI) to evaluate predictive ability and compare methods. The pRKHS method outperformed BayesB, RR-BLUP and BayesA, with pRKHS-E generating the highest correlation of 0.52 (Table 3.5a). Compared to shrinkage methods which used all 1148 markers, pRKHS-E and pRKHS-NE needed only 700 and 100 markers, respectively, to achieve their highest prediction. Furthermore, the optimum CoV for pRKHS-E was 0.3, indicating that few SPC interactions were involved.

A two-year set of experimental data from BarleyCAP was used to measure the predictive ability given an independent set of breeding lines (Table 3.5b). Phenotypes (*i.e.* grain yield (GYD) and plant height (PHT)) and genotypes from Year 2007 were used to fit models and evaluate ten-fold CV performance. The fitted models were then used to predict the phenotype of a different set of 96 lines in Year 2008. For GYD, the pRKHS method substantially outperformed other methods for predicting 2008 phenotypes. pRKHS-NE outperformed other methods using only 1000 markers for both CV assessment and predicted performance. For PHT, pRKHS-NE generated the highest correlation in CV evaluation using 800 markers while pRKHS-E had the highest correlation of EBV and 2008 PHE using 1300 markers. Optimal CoV for pRKHS-E was found at 0.3 in both traits.

3.4 Discussion

This study demonstrates the advantages of using nonparametric methods to estimate true breeding value and to predict phenotypic performance, especially for traits involving epistatic

gene action. The new method is novel because it features a new combination of supervised principal component analysis and reproducing kernel Hilbert spaces, both established statistical methods. The introduction of SPCA complements RKHS by reducing dimensionality and background noise. Two sub-models were constructed to span the range of epistasis involved in trait expression, with pRKHS-E designed to account for low to high epistasis and pRKHS-NE accommodating circumstances in which no or low epistasis exists in the target trait. To evaluate the performance of the pRKHS method, three other shrinkage methods were compared. The results obtained from simulation confirmed that in the absence of epistasis, pRKHS-NE performs more comparably with BayesB than does pRKHS-E (Table 3.1), while pRKHS-E shows better predictive ability when epistasis is present (Tables 3.2, 3.3). In addition, results with actual data indicate that RKHS methods outperform shrinkage methods and can do so with relatively fewer markers (Table 3.5), further confirming the predictive ability of the pRKHS method in real application.

According to selection theory, MBS holds advantage over PS when the genetic correlation (correlation between estimated breeding value and true breeding value) is higher than the correlation of mid-parent and offspring. The consequences that pRKHS outperformed PS in most of the cases (Table 3.1, 3.2, 3.3), its potential use as a means of indirect selection based on marker information alone is highlighted. However, discrepancy of performances from these two methods is expected, which is mainly due to different statistical models constructed for pRKHS-E and pRKHS-NE. In the absence of epistasis, overall underperformance of pRKHS-E (Table 3.1) is mostly attributed to model overfitting. This may be further supported by the results that pRKHS-E had the highest $r_{EBV:PHE}$ but also the lowest $r_{EBV:TBV}$ among five methods in C0 section (Table 3.1), suggesting estimates from pRKHS-E have higher variance and are more biased in

the scenario of no epistasis. As epistasis was increased in simulation scenarios, $r_{EBV:TBV}$ of pRKHS-NE decreased faster than that of pRKHS-E (Table 3.2, Table 3.3), suggesting properly modeling epistasis maintains the advantages of applying MBS.

Note that correlations with pRKHS-E are not overwhelmingly higher compared to pRKHS-NE in low epistasis scenarios (Table 3.2). The result that pRKHS-E outperforms pRKHS-NE in only five out of eight cases indicates pRKHS-NE may function well when a low level of epistasis impacts trait expression. The above phenomenon may be explained by the fact that the optimal CoV for pRKHS-E was 0.3 in low epistasis scenario, wherein about two to three SPC interactions on average were involved in the model (Table 3.4). Overall, 18 SPCs were needed to explain 70% of the variation and were these were included as main effects in the pRKHS-E and pRKHS-NE model (Figure 3.1). Since the principal component score is a linear combination of the weighted marker score, the linear combination of 18 SPC scores may account for a few of SPC interactions. The above argument is further supported by the observations that fitting model using CoV of 0.2 (*i.e.* more SPC interactions) causes multicollinearity in some cases. Overall, features of principal component scores may help the additive model pRKHS-NE fit well in the situation of low epistatic interactions.

Cosine threshold value as mentioned in this study is a nonparametric model diagnostic and used as a criterion to select SPC interactions. As the counterpart of F-statistics in a parametric model (Gu 2002), CoV could theoretically be transformed to a test statistic similar to the F-distribution p-values with some modification (Ma et al. 2009). However, the degrees of freedom for F-distribution which are estimated from the trace of the smoothing matrix change every time a new pair of SPC interactions is fitted. Therefore, the consequential p-value is not monotone with the cosine value, indicating the same cosine value could be assigned for different

p-values in different model fitting, which is misleading to SPC interaction selection, causing loss or false inclusion of interactions. We did some preliminary experimentation by constructing models using a transformed p-value instead of direct CoV for SPC interaction selection and found low predictive ability (data not shown).

In addition to predictive performance, comparisons between pRKHS and shrinkage methods can consider computational load. Several studies (de los Campos et al. 2010; Gianola and van Kaam 2008) have suggested the computational advantages of using nonparametric methods over shrinkage methods. For our models, the cost of the RKHS algorithm is $O\left(nq^2\right)$, where *O* stands for computer power and *n* is sample size and *q* is number of dimensions. With SPCA, *q* is usually around 18 to 20 (Figure 3.1), indicating computational time of pRKHS will be mainly impacted by sample size instead of marker number. With pRKHS, most of the computational load involves constructing reproducing kernels and smoothing matrix and estimating smoothing parameter λ. In contrast, the computation load with Bayesian shrinkage methods is linearly related to the number of features since these are Markov Chain Monte Carlo (MCMC) based, with computational time increasing as the number of number of markers increases. Furthermore, Bayesian methods rely on the convergence of Markov chains to build the posterior distribution, which may require as many as 1000 to 3000 Gibbs sampling iterations during burn-in period.

Model performances were influenced by the underlying genetic architecture of the trait of interest. pRKHS plays an important role when trait expression is influenced by epistasis, whereas shrinkage methods may have higher predictive ability when a trait is controlled by strictly additive gene effects. The genetic architecture represented by BayesB assumes a trait is controlled by a few large gene effects and many small gene effects (Meuwissen *et al.* 2001),

which is similar to the genetic model utilized with pRKHS in this study. Thus, among the three shrinkage methods in this evaluation, BayesB has the most in common with pRKHS with respect to the genetic simulation. Good approximation of the underlying genome seems to contribute to the good performance of BayesB and pRKHS.

The predictive ability of pRKHS is highly related to the included SPCs and their interactions (for pRKHS-E). Bair *et al.* (2006) suggested use of the first several SPCs for prediction and later Li et al. (2011) applied the first three SPCs on genome wide association mapping. With our methods, the number of SPCs to include is flexible and depends on the extent of epistasis; it is quantified by selecting proportion of variation instead of specific numbers. Empirically, we found that with setting a threshold of 70% as the amount of the variation explained by the model and then utilizing only the SPCs associated with that threshold, a good balance between variance explained and goodness-of-fit was achieved in most of the cases through simulation, and this was confirmed with real data applications. However, depending on the crop data and the genetic architecture, the optimal threshold may actually vary by ± 10%, indicating a range of 60% to 80% to achieve best predictability. Cross-validation could be used to find the best number of SPCs to include for a specific data.

Optimal marker density for prediction is a topic of great debate. Some studies advocate use of all markers with dense coverage (Meuwissen et al. 2001), while others found little value in dense coverage of the genome and advocate use of a reduced set of markers for prediction (Long et al. 2007; Luan et al. 2009; VanRaden et al. 2009). Ways of selecting markers also vary and can be based on random selection, genetic distance or LD extent, or entropy reduction, for example. In this study, selection of makers was based on the magnitude of the regression coefficient, *i.e.* the size of the marker effect, and the prediction accuracy is actually increased by

using this criterion. The cost of applying genomic selection on plant breeding is highly related to the number of markers genotyped for each plant. Even with the advent of next generation sequencing and genotyping by sequencing, the potential reduction in genotyping costs makes pRKHS promising in breeding applications, where we will genotype the training population with a dense marker set, but only need to genotype those "significant" markers for future prediction.

Table 3.4 For each scenario with pRKHS, the percent of the total variation explained by top three SPCs (%P1, %P2 and %P3), the number of markers M_{P1}, M_{P2} and M_{P3} included in the respective SPCs, and number of SPC interactions at three given cosine thresholds. Values reflect the lows and highs obtained using various marker subsets (from 500 markers to all markers). Note that larger cosine values are equivalent to smaller p-values.

Scenarios	%P1	%P2	%P3	M_{P1}	M_{P2}	M_{P3}	# of SPC interactions		
							> 0.2	> 0.25	> 0.3
h^2=0.1, E=0	10.4 – 15.1	5.8 – 11.1	5.3 – 9.0	83 – 127	61 – 104	43 – 86	5 – 12	1 – 5	0 – 3
h^2=0.2, E=0	12.2 – 17.7	5.5 – 10.8	5.2 – 8.1	124 – 136	59 – 71	56 – 85	3 – 11	1 – 6	0 – 4
h^2=0.4, E=0	9.3 – 14.9	6.8 – 11.7	5.9 – 9.9	67 – 111	59 – 90	56 – 96	4 – 16	1 – 6	0 – 3
h^2=0.8, E=0	10.4 – 15.3	5.8 – 11.0	5.3 – 9.1	76 – 124	53 – 89	48 – 87	5 – 20	1 – 7	0 – 1
h^2=0.1, E=0.1	11.1 – 17.7	6.1 – 9.7	5.4 – 8.4	105 – 130	55 – 98	50 – 92	4 – 16	1 – 5	0 – 3
h^2=0.2, E=0.1	11.9 – 16.5	5.6 – 11.8	5.1 – 8.2	110 – 125	66 – 85	43 – 88	4 – 12	2 – 7	1 – 4
h^2=0.4, E=0.1	9.2 – 13.7	6.0 – 10.4	5.8 – 9.4	61 – 122	62 – 111	53 – 102	5 – 18	1 – 6	1 – 5
h^2=0.8, E=0.1	11.2 – 13.0	5.6 – 10.6	5.1 – 9.5	69 – 118	54 – 77	44 – 94	6 – 20	2 – 8	1 – 5
h^2=0.1, E=0.5	10.5 – 14.3	5.7 – 9.8	5.1 – 7.8	75 – 120	57 – 86	48 – 103	5 – 18	2 – 7	1 – 2
h^2=0.2, E=0.5	12.0 – 18.7	6.4 – 10.2	5.7 – 7.0	131 – 137	54 – 95	37 – 71	3 – 17	2 – 7	1 – 4
h^2=0.4, E=0.5	12.1 – 18.5	5.5 – 11.7	5.0 – 7.2	122 – 129	83 – 99	41 – 74	5 – 18	3 – 7	1 – 5
h^2=0.8, E=0.5	11.2 – 18.3	5.8 – 10.5	5.1 – 8.6	76 – 126	52 – 107	45 – 96	6 – 21	2 – 9	2 – 5

Table 3.5 Applying pRKHS to real life scenarios, Pearson correlation coefficients between estimated breeding value (EBV) and phenotype obtained from (a) five-fold cross-validation (CV) implemented for maize anthesis-silking interval (ASI) and (b) ten-fold CV using genotypes and phenotypes of barley lines in year 2007 and prediction based on genotypes of different lines in year 2008 implemented for grain yield (GYD) and plant height (PHT) for each of the 5 statistical methods. The number of markers used in each analysis is given, with the optimal number shown for RKHS methods; results were averaged across five repeated fitting. Optimal cosine value was 0.3 for RKHS-E across all datasets.

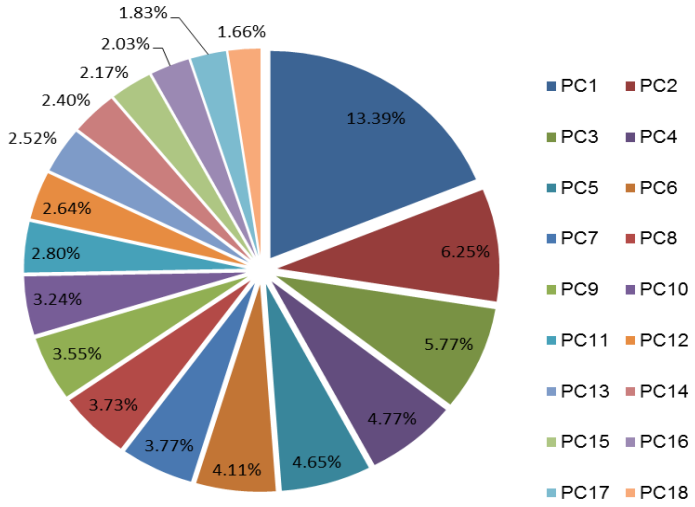
(a)

Trait	CV	Methods	Marker Number	Correlation
ASI	CV	RR-BLUP	1148	0.495
	CV	Bayes-A	1148	0.388
	CV	Bayes-B	1148	0.495
	CV	pRKHS-E	700	0.520
	CV	pRKHS-NE	100	0.515

(b)

Traits	2007 / 2008	Methods	Marker Number	Correlation
GYD	2007	RR-BLUP	1642	0.457
	2007	Bayes-A	1642	0.414
	2007	Bayes-B	1642	0.506
	2007	pRKHS-E	700	0.472
	2007	pRKHS-NE	1000	0.532
	2008	RR-BLUP	1642	0.130
	2008	Bayes-A	1642	0.197
	2008	Bayes-B	1642	0.140
	2008	pRKHS-E	700	0.251
	2008	pRKHS-NE	1000	0.257
PHT	2007	RR-BLUP	1642	0.509
	2007	Bayes-A	1642	0.485
	2007	Bayes-B	1642	0.480
	2007	pRKHS-E	1300	0.495
	2007	pRKHS-NE	800	0.527
	2008	RR-BLUP	1642	-0.065
	2008	Bayes-A	1642	-0.07
	2008	Bayes-B	1642	-0.032
	2008	pRKHS-E	1300	0.179
	2008	pRKHS-NE	800	0.075

Figure 3.1 Mean percentage of variation (across the 12 simulation scenarios) explained by the top 18 SPCs with pRKHS, which together explain 70% of the total variation.



The reduced marker approach used with the new pRKHS method seems to confer some advantages. When no epistasis is present, Bayesian methods perform well with utilization of the full marker information. However, the results that pRKHS-NE had slightly lower prediction accuracy than BayesB (Table 3.1), suggest a near-similar level of predictive ability may be enabled even if partial marker information is used. With pRKHS methods, a 'preselection' procedure is applied before doing PCA to filter out "non-significant" markers. This increases the probability that the subsequent supervised principal components are in good association with the trait of interest (Bair et al. 2006). More importantly, the nonlinearity feature of SPCA which is due to initial marker selection falls into the category of RKHS regression well. Furthermore, PCA serves not only for dimension reduction but also clustering. In simulation, influential markers of each SPC except the first SPC, which contains markers from all ten chromosomes, usually come from one or two linkage groups (chromosomes). Therefore, one SPC is considered to be one or two large haplotypes and the SPC interaction presents the haplotype interactions instead of single marker interaction. Furthermore, methods using haplotypes have been proved to

show higher predictive ability than those only using single marker (Akey et al. 2001; Calus et al. 2008).

Besides prediction, use of pRKHS facilitates inferences about the extent of epistasis involved with a trait of interest. For maize trait ASI with heritability estimated at 0.8 (Buckler et al. 2009), pRKHS-E with optimal cosine value of 0.3 and pRKHS-NE produced comparable results and outperformed RR-BLUP, BayesA and BayesB that only include additive effects (Table 3.5a), indicating that inclusion of a few pairs of SPC interactions in the model increases prediction. The above results not only correspond to the case of simulated low epistasis scenario with $h^2 = 0.8$ (Table 3.2) but also are consistent with the conclusions by Buckler et al. (2009) who suggested that ASI may involve some low level of epistasis . For GYD and PHT in barley, Xu and Jia (2007) concluded that epistasis contributes little to genetic variance for self-pollinated species based on work with a dihaploid population derived from cultivated parents although Von Korff et al. (2010) later found strong epistatic interactions existed in plant height and yield traits in barley and attributed the reason to use of exotic parents and different statistical approaches. As shown in Table 3.5b, our results align with the low epistasis conclusions from Xu and Jia (2007) as pRKHS-E involving a few interactions (cosine value equals 0.3) and pRKHS-NE have comparable predictive ability and both methods are more predictive than the three shrinkage methods.

Overall, the pRKHS method performs well in estimating breeding value and predicting performance when epistasis explains certain proportion of the phenotypic variation. The rate of genetic gain may be enhanced to a certain degree depending on the underlying epistatic extent. Furthermore, pRKHS can be adapted to different types of genetic architectures, *i.e.* epistatic extent and linkage disequilibrium, through tuning CoV and MS, respectively. Compared to other

methods, pRKHS is not only for prediction purposes but also has the capacity to facilitate inferences about the extent of epistasis involved with a trait of interest, which helps scientists to unravel mysteries about the genetic architecture of complex traits. The new nonparametric methods can be readily extended to account for dominance effects and other semi-parametric methods of dealing with some covariates, *e.g.* population structure, typically managed in a parametric manner.

Chapter 4 - Optimization of parameters for successful outcome of version testing in marker-aided trait integration

4.1 Introduction

Since the commercial release of the first transgenic maize hybrids in 1996, breeding efforts have placed greater importance on value-added traits accessed through transformation (Moose and Mumm, 2008). Today, hybrids generally contain multiple transgenic events. SmartStax™, for example, offers 8 different transgenes for insect resistance and herbicide tolerance through incorporation of 4 events (Monsanto News, 2009). By the year 2030, as many as 15-20 biotech traits may be routinely incorporated in new corn hybrids (Fraley 2012).

Marker-aided trait integration (MATI) is the process by which a target hybrid is converted to add the expression of value-added traits to the comprehensive performance package represented by that genotype. The goal is to recover all the attributes of the target hybrid, with the addition of the specified value-added traits. In maize, this process utilizes the backcross method to incorporate events of interest; thus, MATI involves 4 main steps. The first step focuses on introgression of single events into a specific parent of a target hybrid (recurrent parent, RP), with an event defined by the specific DNA added to the host genome by the transformation process and the exact site of this DNA insertion. The second step involves pyramiding of the single events in the RP. The third step focuses on stable expression of the transgenic events by self-pollinating the converted RP to lead to homozygous status of all events. Lastly, the converted target hybrid is formed by hybridizing the converted RP(s) and evaluated in performance testing to ensure the recovery of the target hybrid performance plus the expression of all value-added traits, which is the emphasis of this study.

The activities of first three steps have been the subject by numerous investigations to optimize breeding strategies (Hospital 2001; Hospital and Charcosset 1997; (Hospital et al. 1992) Visscher et al. 1996). In addition, simulation programs such as PLABSIM and Plabsoft (Frisch et al. 2000; Maurer 2008) have been developed to guide breeders' decisions with programmatic (Sun et al. 2011). Nonetheless, little has been published on the topic of 'version testing', which is a key element of the fourth step in MATI as its outcome determines the success or failure of the MATI process. It is defined as the procedure of yield testing several 'versions' of the converted inbred/hybrids, each of which contain all the transgenic events of interest but reflect a unique distribution of non-recurrent parent (NRP) germplasm residual to MATI.

This study builds upon other studies in the Mumm Lab to optimize MATI, namely efforts to identify 'best' strategies for marker-aided introgression. Furthermore, this work also benefits from an update of the genome model to incorporate information relevant to the distribution of genetic effects based on meta-analyses of previous QTL reports (see Chapter 2.3). This upgrade to the genome model reflects training from real data on additive and dominance effects.

The present work reflects a case study involving MATI of 15 events, with the female RP converted for 8 events and the male RP converted for 7 events (Peng et al. unpublished). Through computer simulation, we seek to explore the relationship between parameters that impact the success of the MATI outcome, specifically, minimal number of RP versions required for each parental conversion, the proportion of residual NRP germplasm, and probability of recovering at least 1 hybrid version with performance equivalent to the unconverted target hybrid. Furthermore, we consider ways to minimize the number of hybrid versions to be evaluated. By exploring the potential success rate of the multiple events introgression, we paved

the way for optimizing certain decisions during single event introgression, such as the optimal number of versions to retain and the proportions of the remaining NRP DNA.

4.2 Materials and methods

Background

In this study, we simulated introgression of fifteen events into a target elite hybrid, introgressing 8 and 7 events into the two inbred parents P1 (female RP) and P2 (male RP), respectively. In the process of generating inbred versions on each side of the pedigree, genetic variations among versions were the outcome of the single event introgression stage, where the number of versions of each single event RP conversion was the same as the number of final inbred versions created to proceed through pyramiding and selfing in parallel (Figure 4.1). For example, in case of generating five female versions introgressed with eight events, five versions were retained for each event at the stage of single event introgression.

Genome model

The genome model for simulation was constructed according to the published maize ISU–IBM genetic map, with a total of 1788cM (Fu et al. 2006). Genetic markers were evenly spaced on the chromosome at 0.2cM interval, for a total of 8950 markers across the whole genome. Events to be introgressed were assumed to be inserted into different chromosomes, avoiding issues related to linkage. For example, 8 events introgressed in female RP were arbitrarily placed on chromosomes 1, 3, 4, 5, 7, 8, 9, and 10 and seven events stacked in male RP were randomly positioned on chromosomes 1, 2, 3, 4, 6, 8, and 9. Events were assumed to be traceable by a single marker and no genetic variation was considered for the event expression. Equivalent performance between the converted and unconverted target hybrid was based on grain yield, assumed to be controlled by one hundred QTLs randomly positioned across the genome with magnitude of effects in keeping with the distribution of additive and dominance

effects derived in a previous study (see Chapter 2.3). In the present study, a QTL was assumed to be a gene cluster with five genes per QTL. The inter-genic distance was set to 0.2cM, for a total of 1cM genetic distance to span a QTL. All genes and markers were assumed to be bi-allelic and informative (polymorphic), thus, resulting in a QTL represented by a multi-allelic haplotype. Another assumption in this study was that the alternate alleles were fixed on opposite sides of the pedigree through generations of selections to maximize heterosis; in other words, heterosis was assumed to be purely caused by dominance effects. Before introgression, alleles from male and female side were set in advance to *'G'* and *'g'*, respectively. However, the residual NRP DNA may lead to certain loci with homozygous (*e.g. GG* or *gg*) status instead of heterozygous in the hybrid conversions, especially in cases where the original transformant line originated from the opposite heterotic group from one of the RPs.

The genotypic value for *i*th individual was calculated according to (Cockerham 1954)

$$Geno_i = \sum_{k=1}^{N} u_{ik}\alpha_{1k} + v_{ik}d_k \,, \qquad [1]$$

where design elements u_k and v_k were defined according to the traditional F_∞ model as

$$u_k = \begin{cases} 1 & GG \\ 0 & Gg \\ -1 & gg \end{cases} \quad \text{and} \quad v_k = \begin{cases} 0 & GG \\ 1 & Gg \\ 0 & gg \end{cases},$$

and parameters α_{1k} and d_k were additive and dominance effects of *k*th gene. Additive effects $\alpha_1\left(\alpha_2\right)$ defined as the effect of homozygote carrying the allele from male (female) were drawn from a normal distribution $N\left(0, 0.044\sigma_p^2\right)$, where σ_p was the phenotypic standard deviation of inbred parents. α_{1k} was used in [1] because $u_k = 1$ when male allele '*G*' was in homozygous. And dominance effects, *d*, were obtained from the product of homozygous effects (*a*) defined as half

of the difference between two homozygotes ($a = \frac{\alpha_1 - \alpha_2}{2}$) (Falconer and Mackay 1996) and the dominance coefficients (d/a ratio) which were sampled from a normal distribution $N(0.152, 0.392)$.

The genotypic variance was calculated as follows

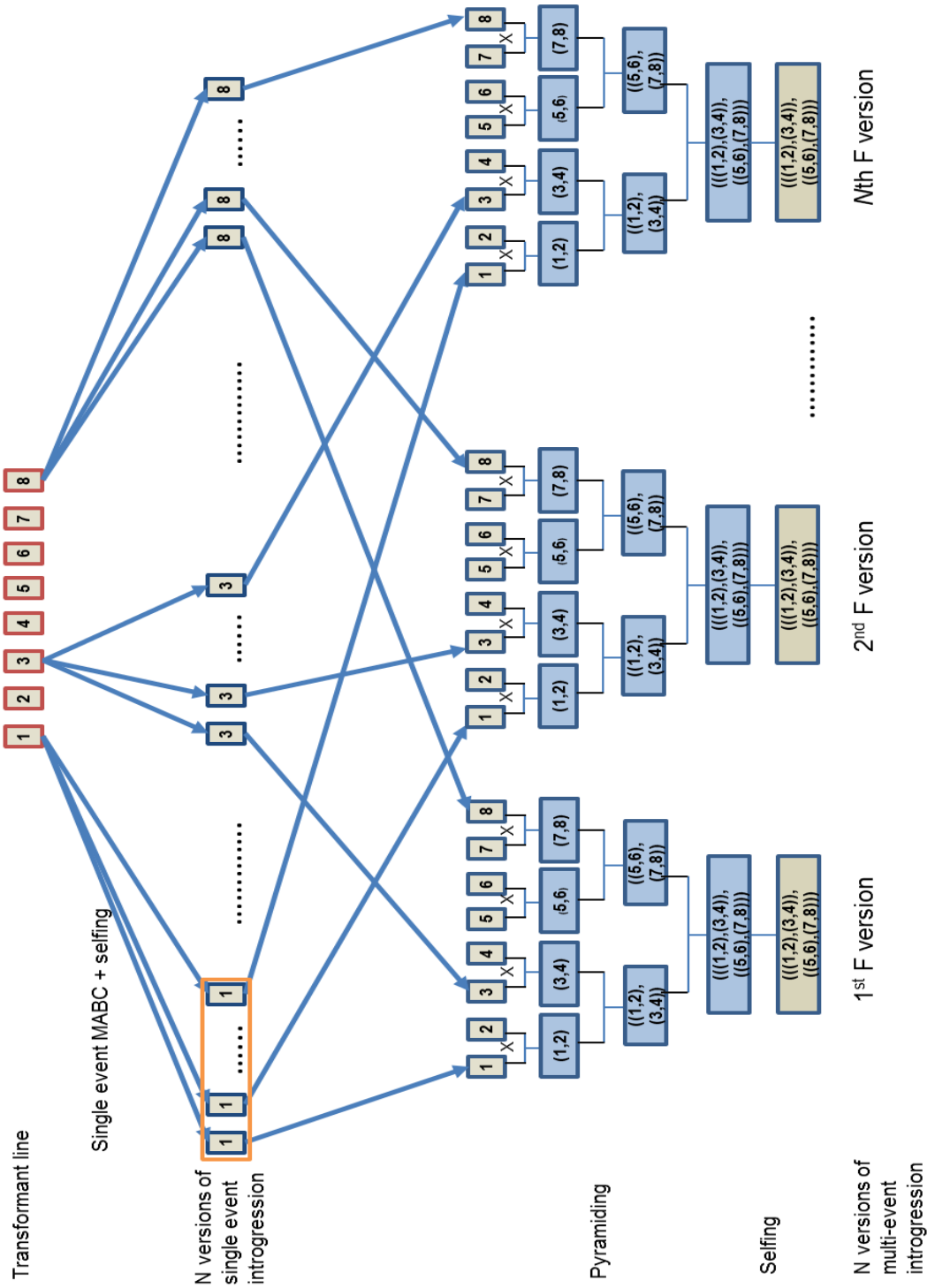
$$V_G = V_A + V_D = \sum_{k=1}^{N} \sigma_{a_k}^2 + \sum_{k=1}^{N} \sigma_{d_k}^2, \qquad [2]$$

where N was the number of gene loci, and $\sigma_{a_k}^2$ and $\sigma_{d_k}^2$ were additive and dominance variance for k^{th} locus. In case of the hybrid population derived from two inbred lines, the allele frequency was 0.5 for all segregating genes and the additive and dominance variance of a single locus became $\sigma_a^2 = \frac{1}{2}a^2$ and $\sigma_d^2 = \frac{1}{4}d^2$ (Falconer and Mackay 1996), leading [2] to

$$V_A = \frac{1}{2} * \sum_{k=1}^{N} a_k^2 \qquad \text{and} \qquad V_D = \frac{1}{4} * \sum_{k}^{N} d_k^2. \qquad [3]$$

Narrow sense heritability of yield was assumed to be 0.4, causing the error variance to be $V_e = 1.5V_A - V_D$. Thus, simulated hybrids had an expected mean (± standard deviation) of 235.6 (± 48.1) bu/ac (see Appendix A). Since dominance effects were absent in inbred lines, breeding values of RP versions were equivalent to their genotypic values, which had an expected value of zero.

Figure 4.1 Illustration of the process generating female inbred versions stacked with eight events. Only one line was kept during pyramiding and selfing stage. Yellow and blue box indicate homozygous and heterozygous state of the event, respectively. Red and blue broader indicate different genetic background.



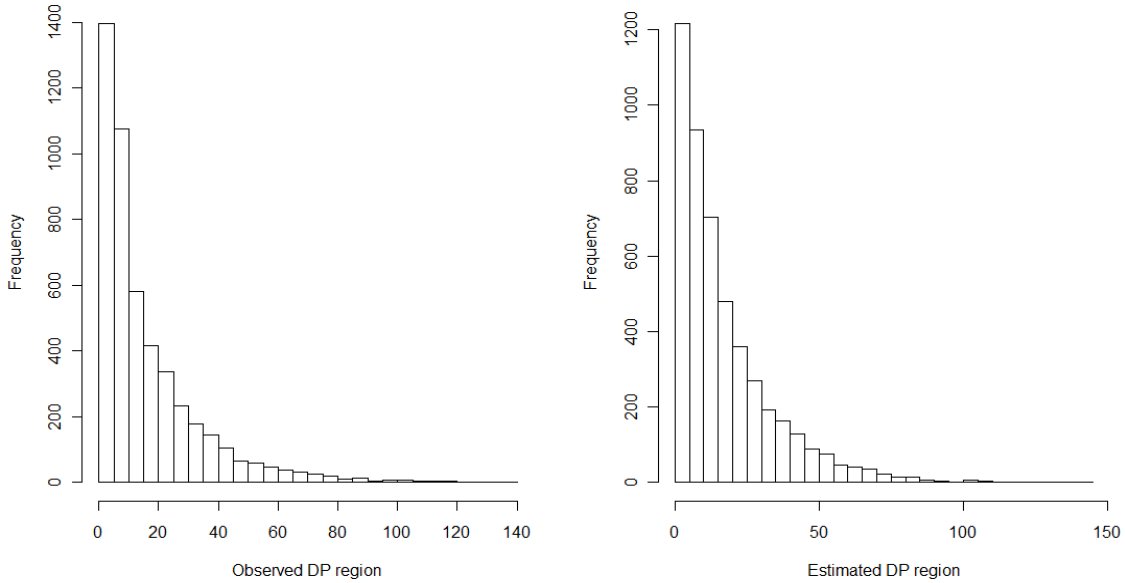
Donor parent genome after introgression

The length of residual NRP germplasm segments at the close of single event conversion of RPs, expressed in terms of units of genetic distance (cM), was assumed to follow exponential distribution (Figure 4.2). Since the smallest NRP segment was always larger than zero, we fitted a truncated exponential distribution to the data. The density for the truncated distribution was

$$P(y_i \mid \lambda) = \frac{\lambda e^{-\lambda y_i}}{\int_c^{\infty} \lambda e^{-\lambda y} dy} = \lambda e^{-\lambda(y_i - c)}, \qquad [4]$$

where y_i was i^{th} observed data, c was the truncation point, and λ was the rate. Referring to the previous introgression results (Ting et al. unpublished data), the minimum size of NRP segments in the 20cM flanking regions (FR) bracketing events was 1cM, *i.e.* c = 1. To draw sample Y from the truncated distribution, we simply simulated $X \sim EXP(\hat{\lambda})$, where $\hat{\lambda}$ was obtained by taking the reciprocal of expected genetic length of non-flanking regions (NFR) defined as NRP residing outside of flanking regions of the event and further took $Y = X + 1$.

Figure 4.2 Histograms of observed and estimated remaining donor parent regions (cM). The estimated values were obtained from an exponential distribution truncated at 1cM.



To represent pyramiding of the single events in a specific RP, we repeated the above sampling procedure over n_e times where n_e was the total number of events to be introgressed. The above procedure was based on assumptions that the introgression of one event was independent of others and that the pyramiding process contributed little to genetic content shuffling. When introgressing 15 events into the target hybrid, with 8 events to a female line and 7 events to a male line, the total amount of residual NRP in a hybrid conversion was set to a series of values: 180, 160, 140, and 120cM, which corresponds to 10%, 8.9%, 7.8%, and 6.7% proportion on NRP germplasm across the genome, respectively. Thus, for each event in the converted target hybrid, an average of 12, 10.6, 9.3, and 8cM, respectively, was associated with each single event conversion, 1cM of which was located in the FR and the rest located at other genomic locations across the genome. The 1cM-length FR fragments were randomly positioned in the flanking regions of the events in the simulation whereas the balances of the NRP fragments randomly placed across the whole genome were sampled from [4]. In *silico* female (male) inbred versions were thus generated by repeating the above sampling-positioning procedure $n_f(n_m)$ times, where $n_f(n_m)$ was the number of female (male) versions. The analysis considered $n_f(n_m) \leq 5$ to find the optimal number of inbred versions on each pedigree side to stay within realistic bounds.

For each conversion of a target hybrid, a total of $n_f * n_m$ versions were derived from the cross between various versions of conversions of female and male RPs. Two scenarios were implemented to calculate the success rate (SR) of finding at least one hybrid conversion with equivalent performance of the unconverted target hybrid (UTH). 1) Full-hybrid scenario: all hybrid conversions (ranging from 9 to 25 which relied on the number of versions selected to cross) were tested in yield trial and compared to the UTH performance. And 2) partial-hybrid

scenario: several inbred versions from each side of the pedigree were picked to form at most 9 hybrid conversions to perform yield testing. Selection of converted RPs was done in two ways: 1) random selection and 2) selection based on an estimate of breeding value. This simulation scenario (including creating genetic map, positioning QTLs, generating genetic effects, and generating and selecting inbred versions) was repeated 1000 times to calculate SR, that is, the likelihood of covering at least one hybrid conversion with equivalent performance within 3% of the UTH performance based on a specified probability level. To calculate the mean and standard errors of SR, the whole procedure was also repeated six times.

4.3 Results

Simulation results indicated a high degree of relationship between the SR, amount of residual NRP germplasm, and number of versions of each converted RP of the target hybrid. With the increase of NRP germplasm in the genome, the SR of recovering at least one hybrid conversion with equivalent performance decreased. At the NRP level of 120cM, three SRs exceeded the 95% probability level, with 4 versions of the female RP conversion and 5 of the male RP, with 5 versions of the female RP and 4 of male RP, and with 5 versions of each RP. At NRP levels from 140 – 180 cM, the SR exceeded 95% only with 5 versions of each RP (Table 4.1). In general, the SR was positively correlated with the number of versions of the RPs. In each of these cases, the SR assumes that all possible hybrid combinations of RP versions would be evaluated in yield performance trials. In the full-hybrid scenario, around 8% SR rise was observed when the number of hybrid increased from 9 to 25 (Table 4.1).

Using breeding value to pare down the number of versions to advance to yield trials, the SR associated with testing a subset of all possible hybrid combinations of RP versions was

estimated. In general, approximately 2% SR reduction was realized for testing a subset of 9 hybrids created from the 'best' 3 versions of each RP (Figure 4.2).

The correlation between SR using all and partial hybrid conversions was associated with the number of hybrid conversions involved in the partial-hybrid scenario and represented by the ratio between SR derived from partial and all hybrid conversions. The higher the ratio is, the larger the probability that partial hybrid conversions could contain the hybrid that recover UTH. As shown in Table 4.3, around 0.02 of the ratio rise was gained when two more hybrid conversions to be tested. Compared to employing two male and female versions which lead to a total of four hybrid conversions, usage of nine hybrid conversions derived from three male and female versions increased the ratio by nearly 0.06. No significant difference was observed between cases of using the same number of hybrid but different inbred versions. Meanwhile, usage of breeding value to select inbred versions had higher efficiency than random selection did. The best partial-hybrid scenario was found by choosing three out of five versions from both sides, making a total of nine hybrid conversions (Table 4.2, 4.3).

4.4 Discussion

The value of computer simulation in guiding critical decisions facing the plant breeder is enhanced if the underlying models such as the genome model accurately portray real genetic processes and true genetic architecture. In the present study, we used distributions of additive and dominance effects which were derived from meta-analyses of previously published QTL studies to update the genome model used to simulate genetic effects. This aided us in estimating the positional effects of NRP germplasm remaining in finalized conversions relative to recovery of yield performance of the converted target hybrid introgressed with 15 events. One difficulty of the research was to simulate heterosis which was mainly endorsed by two important theories

such as dominance and overdominance (Birchler et al. 2003). In this study, we adopted the suggestions from Edwards et al. (1987) who attributed the QTL overdominance effects detected in corn yield to the repulsive linkage of several genes with partial dominance. In detail, additive effects of genes sampled from the normal distribution with null mean were automatically assigned positive or negative sign, indicating random repulsive linkages between genes. And the positive mean of dominance coefficients, *i.e.* 0.152, indicated the overall partial dominance gene actions. The results of small additive effects and overall partial dominance gene actions conformed to previous conclusions that yield was composed of large number of small effect genes (Stuber et al. 1987) and dominance coefficients tend to have positive direction (Kacser and Burns 1981). These settings fit the expected hybrid mean and standard deviation in a reasonable range (Appendix A).

Besides heterosis, it is also very important to have an accurate way of simulating remaining NRP segments residual to MATI. Because the overall amount and locations of the NRP fragments is influenced by recombination and selection during MATI, we directly utilized the results obtained by Peng *et al.* (unpublished) on the optimal MATI strategy for single event introgression. As shown in Figure 4.2, the simulated NRP results were fitted well by a truncated exponential distribution, suggesting the optimal MATI strategy had been incorporated into the simulation by sampling NRP fragments from the distribution.

Table 4.1 Estimated SR of recovering ≥1 hybrid conversion with yield within 3% of the unconverted target hybrid given performance testing of all possible hybrid combinations of various number of versions of RP conversions. Values were obtained from 1000 simulations with six repeats. Standard errors of the estimates were 0.56-0.59%.

%NRP	NRP (cM)	Single event conversion (cM)	# female versions	# male versions		
				3	4	5
6.7%	120	8	3	88.82%	92.23%	93.68%
			4	92.20%	93.73%	95.05%
			5	93.25%	95.20%	96.60%
7.8%	140	9.3	3	88.73%	91.93%	92.58%
			4	90.82%	93.13%	94.67%
			5	93.05%	94.88%	95.87%
8.9%	160	10.7	3	88.15%	90.85%	91.92%
			4	90.70%	93.02%	94.45%
			5	92.30%	94.73%	95.32%
10%	180	12	3	87.13%	90.73%	91.18%
			4	90.50%	92.57%	93.75%
			5	92.03%	93.97%	95.18%

Table 4.2 Estimated SR of recovering ≥1 hybrid conversion with yield within 3% of the unconverted target hybrid given performance testing of 9 hybrid combinations of various number of versions of RP conversions after selecting the 'best' 3 versions of each RP from the total number of versions created. Values were obtained from 1000 simulations with six repeats. The standard errors of estimates were 0.51-0.57%.

NRP (cM)	Total # of female versions	Total # of male versions		
		3	4	5
120	3	88.82%	90.45%	91.62%
	4	90.42%	91.67%	92.96%
	5	91.20%	93.11%	94.47%
140	3	88.73%	90.16%	90.55%
	4	89.06%	91.08%	92.58%
	5	91.00%	92.80%	93.76%
160	3	88.15%	89.10%	89.89%
	4	88.95%	90.97%	92.37%
	5	90.27%	92.65%	93.22%
180	3	87.13%	88.98%	89.18%
	4	88.75%	90.53%	91.69%
	5	90.01%	91.90%	93.09%

Selecting the 'best' 3 versions from each set of RP conversions using breeding value to identify individuals with good genetic potential increased the success rate of recovering UTH performance. For example, Ødegård et al. (2009) also incorporated genomic selection into marker-assisted introgression to track both the gene being introgressed and total genetic merit, especially for low heritability traits like yield. Although the SR increased as the number of inbred versions increased, it may not be necessary to perform yield testing using all hybrid conversions. As shown, only a 2% deduction in SR was observed when switching from the scenario of testing all hybrid conversions to that of testing only nine hybrids selected by using breeding value criteria. Thus, the comparison of the partial-hybrid scenario to the full-hybrid scenario with ≥95% SR emphasizes resource allocation: 4+5 versions of RPs produced plus 20

hybrids yield tested broadly versus 5+5 versions of RPs produced plus 9 hybrids yield tested broadly (Tables 4.1 and 4.2).

In the present simulation study, we used true breeding value (TBV) to serve as the selection index, whereas in real life where TBV is unknown. The estimated breeding value (EBV) predicted by dense genetic markers could be used as an alternative but might have slightly lower precision due to low heritability of yield. Furthermore, we didn't take into account dominance effect for inbred version selection. However, it might be useful to perform direct selection on hybrid conversions by predicting single cross performances using genomic information, in which case, specific combining ability as well as general combining ability could be estimated (Arbelbide et al. 2006).

In short, version testing is implemented in a trait integration program to identify at least one conversion of a target hybrid that satisfies simultaneously the requirements of incorporating all the transgenic events of interest and recovering UTH performance. Unless this requirement is met, the MATI does not achieve commercial reality and all efforts have been in vain. We used computer simulation to explore and develop some guidelines for such a program.

Table 4.3 The ratio between the success rates derived from using partial and full hybrid conversions. Partial hybrid conversions were derived from crossing between inbred versions selected by 1) randomly picking or 2) using breeding value (BV) criteria. Standard errors of the estimates were 0.38-0.51%. Large ratio value indicates high similarity between lines involved in partial and full hybrid conversions.

# female version	# male versions	Random	BV
2	2	0.81	0.92
2	3	0.85	0.95
3	2	0.85	0.95
2	4	0.86	0.96
4	2	0.86	0.96
3	3	0.89	0.98

Appendix A

Hybrid mean

The target hybrid performance (*H*) was the sum of mid-parent value and heterosis which was assumed to be purely caused by dominance effects. Thus, expected target hybrid performance $E(H) = \mu_p + N * E(d)$, where μ_p was the mean of two inbred parents, *N* was the number of gene locus and $E(d)$ was the mean of dominance effect. Given the distribution of additive effects and dominance coefficients, $E(d) = E\left(a * \frac{d}{a}\right) = E(a)E\left(\frac{d}{a}\right) = 0.152 * E(a)$, where *a* was the homozygous effect following a truncated distribution of additive effects, *i.e.* $a \sim TN\left(0, 0.044 * \sigma_p^2\right)$, with truncation point at 0. And $E(a) = \int_0^\infty x f(x) = \int_0^\infty \frac{2x}{\sqrt{2\pi}\sigma} e^{\frac{-x^2}{2\sigma^2}} dx = \sigma\sqrt{\frac{2}{\pi}}$, where $f(x)$ was the probability density function of *TN* and σ was the standard deviation of additive effects, *i.e.* $\sigma = \sqrt{0.044}\sigma_p$. Mean and standard deviation of inbred population, μ_p and σ_p were set to 70 bu/ac and 13 bu/ac, respectively. Both values were derived from the yield performance of 12 elite inbred lines, representing important heterotic sub-groups in US maize commercial germplasm (Unpublished data from Brian Mansfield). Given the above values, the expected hybrid was $E(H) = \mu_p + N * E(d) = 70 + 500 * 0.152 * \sqrt{\frac{2}{\pi}}\sigma = 235.6$ bu/ac.

Hybrid standard deviation

Based on [3], the expected additive and dominance were $E(V_A) = 0.5 * 500 * E(a^2)$, and $E(V_D) = 0.25 * 500 * E(d^2) = 125 * E(a^2)E\left(\left(\frac{d}{a}\right)^2\right)$, where $E(a^2) = E\left(\left(\frac{\alpha_1 - \alpha_2}{2}\right)^2\right) = \frac{1}{2}E(\alpha_1^2)$

(Falconer and Mackay 1996). Therefore, $E(V_A) = 125 * 0.044 * 13^2 = 925.5$, and

$E(V_D) = 125 * \frac{0.044 * 13^2}{2} * 0.392 = 182.2$. Given narrow sense heritability of 0.4, the expected error variance was $E(V_e) = 1.5E(V_A) - E(V_D) = 1212.1$. And the total phenotypic standard deviation of the hybrid was $\sqrt{V_A + V_D + V_e}$ = 48.1 bu/ac.

References

Akey J, Jin L, Xiong M (2001) Haplotypes vs single marker linkage disequilibrium tests: What do we gain? Eur J Hum Genet 9:291-300

Aldous D (1985) Exchangeability and related topics in *l'école d'été de probabilités de saint-flour, xiii-1983*, Springer, Berlin, pp. 1-198.

Anderson LK, Doyle GG, Brigham B, Carter J, Hooker KD, Lai A, Rice M, Stack SM (2003) High-resolution crossover maps for each bivalent of *zea mays* using recombination nodules. Genetics 165:849-865

Austin DF, Lee M, Veldboom LR, Hallauer AR (2000) Genetic mapping in maize with hybrid progeny across testers and generations: grain yield and grain moisture. Crop Sci 40:30-39

Bair E, Hastie T, Paul D, Tibshirani R (2006) Prediction by supervised principal components. J Am Stat Assoc 101:119-137

Barchi M, Roig I, Di Giacomo M, de Rooij DG, Keeney S, Jasin M (2008) ATM promotes the obligate XY crossover and both crossover control and chromosome axis integrity on autosomes. PLoS Genet 4:e1000076

Beavis WD (1998) QTL analyses: Power, precision, and accuracy. In: Paterson AH (ed) Molecular dissection of complex traits. CRC Press, New York, pp 145-162

Bennewitz J, Meuwissen THE (2010) The distribution of QTL additive and dominance effects in porcine F2 crosses. J Anim Breed Genet 127:171-179

Bennewitz J, Solberg T, Meuwissen T (2009) Genomic breeding value estimation using nonparametric additive regression models. Genet Sel Evol 41:20

Bernardo R (1990) Identifying populations useful for improving parents of a single cross based on net transfer of alleles. Theor Appl Genet 80:349-352

Bernardo R (1993) Estimation of coefficient of coancestry using molecular markers in maize. Theor Appl Genet 85:1055-1062

Bernardo R (1994) Prediction of maize single-cross performance using RFLPs and information from related hybrids. Crop Sci 34:20-25

Bernardo R (2002) Breeding for quantitative traits in plants. Stemma Press

Bernardo R (2009) Genomewide selection for rapid introgression of exotic germplasm in maize. Crop Sci 49:419-425

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci 47:1082-1090

Birchler JA, Auger DL, Riddle NC (2003) In search of the molecular basis of heterosis. The Plant Cell Online 15:2236-2239

Blanc G, Charcosset A, Mangin B, Gallais A, Moreau L (2006) Connected populations for detecting quantitative trait loci and testing for epistasis: An application in maize. Theor Appl Genet 113:206-224

Bost B, de Vienne D, Hospital F, Moreau L, Dillmann C (2001) Genetic and nongenetic bases for the L-shaped distribution of quantitative trait loci effects. Genetics 157:1773-1787

Bost B, Dillmann C, de Vienne D (1999) Fluxes and metabolic pools as model traits for quantitative genetics. I. The L-shaped distribution of gene effects. Genetics 153:2001-2012

Briggs WH, McMullen MD, Gaut BS, Doebley J (2007) Linkage mapping of domestication loci in a large maize–teosinte backcross resource. Genetics 177:1915-1928

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC (2009) The genetic architecture of maize flowering time. Science 325:714-718

Bulmer MG (1985) The mathematical theory of quantitative genetics. Oxford University Press, Oxford

Burkhamer RL, Lanning SP, Martens RJ, Martin JM, Talbert LE (1998) Predicting progeny variance from parental divergence in hard red spring wheat. Crop Sci 38:243-248

Burrows PM (1975) Expected selection differentials for directional selection. Biometrics 28:1091-1100

Byrd RH, Lu P, Nocedal J, Zhu C (1995) A limited memory algorithm for bound constrained optimization. SIAM J Scientific Computing 16:1190-1208

Caldwell KS, Russell J, Langridge P, Powell W (2006) Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *hordeum vulgare*. Genetics 172:557-567

Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF (2008) Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553-561

Calus MPL, Veerkamp RF (2007) Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. J Anim Breed Genet 124:362-368

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963-971

Coburn JR, Temnykh SV, Paul EM, McCouch SR (2002) Design and application of microsatellite marker panels for semiautomated genotyping of rice (*oryza sativa l.*). Crop Sci 42:2092-2099

Cockerham CC (1954) An extension of the concept of partitioning hereditary variacne for analysis of covariances among relatives when epistasis is present. Genetics 39:859-882

Cooper M, Podlich DW (2002) The E(NK) model: Extending the NK model to incorporate gene-by-environment interactions and epistasis for diploid genomes. Complexity 7:31-47

Cooper M, Podlich DW, Luo L (2007) Modeling QTL effects and MAS in plant breeding. Genomics-assisted crop improvement, pp 57-95

Copenhaver GP, Housworth EA, Stahl FW (2002) Crossover interference in arabidopsis. Genetics 160:1631-1639

Crosby JL (1973) Computer simulation in genetics. Wiley, Hoboken, NJ

Crossa J, de los Campos G, Perez P, Gianola D, Burgueno J, Araus JL, Makumbi D, Singh R, Dreisigacker S, Yan J, Arief V, Banziger M, Braun H-J (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713-724

Crow JF, Kimura M (2009) An introduction to population genetics theory. Blackburn Press, Caldwell, NJ

de los Campos G, Gianola D, Rosa JMG, Weigel AK, Crossa J (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res 92:295-308

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, Weigel K, Cotes JM (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigrees. Genetics 182:375-385

Dintinger J, Verger D, Caiveau S, Risterucci AM, Gilles J, Chiroleu F, Courtois B, Reynaud B, Hamon P (2005) Genetic mapping of maize stripe disease resistance from the mascarene source. Theor Appl Genet 111:347-359
Dudley JW (1982) Theory for transfer of alleles. Crop Sci 22:631-637
Dudley JW (1984) A method of identifying lines for use in improving parents of a single cross. Crop Sci 24:355-357
Dudley JW (1987) Modification of methods for identifying populations to be used for improving parents of elite single crosses. Crop Sci 27:940-943
Dudley JW (2004) Breeding: Choice of parents. In: Goodman RM (ed) Encyclopedia of plant and crop science. Taylor & Francis, London., pp 215-217
Dudley JW, Johnson GR (2009) Epistatic models improve prediction of performance in corn. Crop Sci 49:1533-1533
Dudley JW, Johnson GR (2010) Epistatic models improve between year prediction and prediction of testcross performance in corn. Crop Sci 50:763-769
Dudley JW, Maroof MAS, Rufener GK (1992) Molecular marker information and selection of parents in corn breeding programs. Crop Sci 32:301-304
Eathington SR, Crosbie TM, Edwards MD, Reiter RS, Bull JK (2007) Molecular markers in a commercial breeding program. Crop Sci 47:154-163
Edwards MD, Stuber CW, Wendel JF (1987) Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics 116:113-125
Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics. Longman and Company, Essex, UK
Falque M, Anderson LK, Stack SM, Gauthier F, Martin OC (2009) Two types of meiotic crossovers coexist in maize. Plant Cell 21:3915-3925
Fan J (1996) Local Polynomial Modelling and Its Applications. Chapman & Hall/CRC, Boca Raton, Florida
Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. Ann Statistics 1:209-230
Fernando R, Grossman M (1989) Marker assisted selection using best linear unbiased prediction. Genet Sel Evol 21:467-477
Foss E, Lande R, Stahl F, Steinberg C (1993) Chiasma interference as a function of genetic distance. Genetics 133:681-691
Fraley, R. Monsanto Company, Food & Agricultural Communications: The Next Frontier, University of Illinois Agricultural Communications, February 17, 2012
Frisch M, Bohn M, Melchinger AE (2000) Computer note. PLABSIM: Software for simulation of marker-assisted backcrossing. J Hered 91:86-87
Frisch M, Thiemann A, Fu J, Schrag T, Scholten S, Melchinger AE (2010) Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. Theor Appl Genet 120:441-450
Fu Y, Wen T-J, Ronin YI, Chen HD, Guo L, Mester DI, Yang Y, Lee M, Korol AB, Ashlock DA, Schnable PS (2006) Genetic dissection of intermated recombinant inbred lines using a new genetic map of maize. Genetics 174:1671-1683
Gao H, Williamson S, Bustamante CD (2007) A Markov Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics 176:1635-1651

Gianola D (2009) Additive genetic variability and the bayesian alphabet. Genetics 183:347
Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761-1776
Gianola D, van Kaam JBCHM (2008) Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178:2289-2303
Gonzalez-Recio O, Gianola D, Long N, Weigel KA, Rosa GJM, Avendano S (2008) Nonparametric methods for incorporating genomic information into genetic evaluations: An application to mortality in broilers. Genetics 178:2305-2313
Gordillo GA, Geiger HH (2008a) Alternative recurrent selection strategies using doubled haploid lines in hybrid maize breeding. Crop Sci 48:911-922
Gordillo GA, Geiger HH (2008b) Mbp (version 1.0): A software package to optimize maize breeding procedures based on doubled haploid lines. J Hered 99:227-231
Grapes L, Dekkers JCM, Rothschild MF, Fernando RL (2004) Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. Genetics 166:1561-1570
Griffing B (1956) A generalized treatment of the use of diallel crosses in quantitative inheritance. Heredity 10:31-50
Gruber MHJ (1998) Improving Efficiency by Shrinkage. Marcel Dekker, New York
Gu C (2002) Smoothing Spline ANOVA Models. Springer-Verlag, New York
Habier D, Fernando R, Kizilkaya K, Garrick D (2011) Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12:186
Haldane JBS (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 8:299-309
Hallauer AR, Pandey S (2006) Plant breeding: The Arnel R. Hallauer international symposium: Chapter 4 defining and achieving plant-breeding goals. Blackwell Publishing, Oxford
Hamblin MT, Salas Fernandez MG, Casa AM, Mitchell SE, Paterson AH, Kresovich S (2005) Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass sorghum bicolor. Genetics 171:1247-1256
Hayes BJ, Goddard ME (2001) The distribution of the effects of genes affecting quantitative traits in livestock. Genet Sel Evol 33:209-229
Heckenberger M, Maurer HP, Melchinger AE, Frisch M (2008) The Plabsoft database: A comprehensive database management system for integrating phenotypic and genomic data in academic and commercial plant breeding programs. Euphytica 161:173-179
Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. Crop Sci 49:1-12
Henderson CR (1984) Applications of Linear Models in Animal Breeding. Can. Catal. Publ. Data, University of Guelph, Canada
Hessel DA, Lawrence CJ, Lauter N (2010) P. 28-39. *In* proc. Cogenfito: A composite genotype finder tool for optimizing isoline selection in maize breeding schemes. Corn Breeders School, 46st, Univ of Illinois Urbana-Champaign, IL
Holland JB (2007) Genetic architecture of complex traits in plants. Curr Opin Plant Biol 10:156-161
Hospital F (2001) Size of donor chromosome segments around introgressed loci and reduction of linkage drag in marker-assisted backcross programs. Genetics 158:1363-1379
Hospital F, Charcosset A (1997) Marker-assisted introgression of quantitative trait loci. Genetics 147:1469-1485

Hospital F, Chevalet C, Mulsant P (1992) Using markers in gene introgression breeding programs. Genetics 132:1199-1210
Housworth EA, Stahl FW (2009) Is there variation in crossover interference levels among chromosomes from human males? Genetics 183:403-405
Huelsenbeck JP, Andolfatto P (2007) Inference of population structure under a Dirichlet Process model. Genetics 175:1787-1802
Hyten DL, Choi I-Y, Song Q, Shoemaker RC, Nelson RL, Costa JM, Specht JE, Cregan PB (2007) Highly variable patterns of linkage disequilibrium in multiple soybean populations. Genetics 175:1937-1944
Ihaka R, Gentleman R (1996) A language for data analysis and graphics. J Comput Graph Stat 5:299-314
Ishii T, Yonezawa K (2007) Optimization of the marker-based procedures for pyramiding genes from multiple donor lines: II. Strategies for selecting the objective homozygous plant. Crop Sci 47:1878-1886
Kacser H, Burns JA (1981) The molecular basis of dominance. Genetics 97:639-666
Karlin KS, Liberman UL (1978) Classifications and comparisons of multilocus recombination distributions. Proc Natl Acad Sci USA 75:332-336
Kearsey MJ, Farquhar AGL (1998) QTL analysis in plants; where are we now? Heredity 80:137-142
Kharkwal MC, Roy D (2004) Plant Breeding - Mendelian to Molecular Approaches: 2. A century of advances in plant breeding methodologies. Narosa Publishing House, New Delhi, India
Kosambi DD (1944) The estimation of map distance from recombination values. Annals of Eugenics 12:172-175
Kuchel H, Ye G, Fox R, Jefferies S (2005) Genetic and economic analysis of a targeted marker-assisted wheat breeding strategy. Mol Breeding 16:67-78
Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743-756
Laurie CC, Chasalow SD, LeDeaux JR, McCarroll R, Bush D, Hauge B, Lai C, Clark D, Rocheford TR, Dudley JW (2004) The genetic architecture of response to long-term artificial selection for oil concentration in the maize kernel. Genetics 168:2141-2155
Lawrence CJ, Schaeffer ML, Seigfried TE, Campbell DA, Harper LC (2007) MaizeGDB's new data types, resources and activities. Nucl Acids Res 35:895-900
Li J, Das K, Fu G, Li R, Wu R (2011) The Bayesian Lasso for genome-wide association studies. Bioinformatics 27:516-523
Long N, Gianola D, Rosa GJM, Weigel KA, Avendaño S (2007) Machine learning classification procedure for selecting snps in genomic selection: Application to early mortality in broilers. J Anim Breed Genet 124:377-389
Longin C, Utz H, Reif JC, Schipprack W, Melchinger AE (2006) Hybrid maize breeding with doubled haploids: I. One-stage versus two-stage selection for testcross performance. Theor Appl Genet 112:903-912
Luan T, Woolliams JA, Lien S, Kent M, Svendsen M, Meuwissen THE (2009) The accuracy of genomic selection in norwegian red cattle assessed by cross validation. Genetics 183:1119-1126
Ma P, Zhong WX, Liu JS (2009) Identifying differentially expressed genes in time course microarray data. Stat in Biosciences 1:144-159

Macciotta N, Gaspa G, Steri R, Pieramati C, Carnier P, Dimauro C (2009) Pre-selection of most significant SNPs for the estimation of genomic breeding values. BMC Proc 3:S14
Macciotta NPP, Gaspa G, Steri R, Nicolazzi EL, Dimauro C, Pieramati C, Cappio-Borlino A (2010) Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. J Dairy Sci 93:2765-2774
Mackay TFC (2001) The genetic architecture of quantitative traits. Annu Rev Genet 35:303-339
Mallick BK, Ghosh D, Ghosh M (2005) Bayesian classification of tumours by using gene expression data. J Roy Stat Soc Ser B (Statistical Methodology) 67:219-234
Malysheva-Otto LV, Ganal MW, Roder MS (2006) Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*hordeum vulgare l.*). BMC Genet 7:6
Mather K (1935) Reduction and equational separation of the chromosomes in bivalents and multivalents. J Genet 30:53-78
Maurer HP (2008) Development and applications of Plabsoft: A computer program for population genetic data analyses and simulations in plant breeding. Department of Applied Genetics and plant breeding. University of Hohenheim, p 59
Maurer HP, Melchinger AE, Frisch M (2007) An incomplete enumeration algorithm for an exact test of Hardy-Weinberg proportions with multiple alleles. Theor Appl Genet 115:193-398
Maurer HP, Melchinger AE, Frisch M (2008) Population genetic simulation and data analysis with plabsoft. Euphytica 161:133-139
McPeek M, Speed T (1995) Modeling interference in genetic recombination. Genetics 139:1031-1044
Melchinger AE, Schmidt W, Geiger HH (1988) Comparison of testcrosses produced from F_2 and first backcross populations in maize. Crop Sci 28:743-749
Messmer R, Fracheboud Y, Bänziger M, Vargas M, Stamp P, Ribaut J-M (2009) Drought stress and tropical maize: QTL-by-environment interactions and stability of QTLs across environments for yield components and secondary traits. Theor Appl Genet 119:913-930
Metz G (1994) Probability of net gain of favorable alleles for improving an elite single cross. Crop Sci 34:668-672
Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829
Mézard C, Vignard J, Drouaud J, Mercier R (2007) The road to crossovers: Plants have their say. Trends in Genetics 23:91-99
Mihaljevic R, Utz HF, Melchinger AE (2005) No evidence for epistasis in hybrid and *per se* performance of elite European flint maize inbreds from generation means and QTL analyses. Crop Sci 45:2605-2613
Moose SP, Mumm RH (2008) Molecular plant breeding as the foundation for 21st century crop improvement. Plant Physiol 147:969-977
Monsanto News (2009) Monsanto, Dow AgroSciences complete U.S. and Canadian regulatory authorizations for SmartStax corn; plans set to launch seed platform on 3 million to 4 million-plus acres. Monsanto News Releases July 20, 2009 < http://monsanto.mediaroom.com/index.php?s=43&item=729 >
Mumm RH (2007) Backcross versus forward breeding in the development of transgenic maize hybrids: Theory and practice. Crop Sci 47:S-164-171
Nadaraya E (1964) On estimating regression. Theor Probab Appl 9:141-142

Neal RM (2000) Markov Chain sampling methods for Dirichlet Process mixture models. J Comput Graph Stat 9:249-265
Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321-3323
Nei M, Li WH (1979) Mathematical models for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5268-5371
Ødegård J, Yazdi MH, Sonesson AK, Meuwissen THE (2009) Incorporating desirable genetic characteristics from an inferior into a superior population using genomic selection. Genetics 181:737-745
Omori F, Mano Y (2007) QTL mapping of root angle in F2 populations from maize 'B73' x teosinte '*zea luxurians*' Plant Root 1:57-65
Otto SP, Jones CD (2000) Detecting the undetected: Estimating the total number of loci underlying a quantitative trait. Genetics 156:2093-2107
Panter DM, Allen FL (1995) Using best linear unbiased predictions to enhance breeding for yield in soybean: I. Choosing parents. Crop Sci 35:397-405
Piepho HP (2009) Ridge regression and extensions for genomewide selection in maize. Crop Sci 49:1165-1176
Plummer M, Best N, Cowles K, Vines K (2006) Coda: Output analysis and diagnostics for mcmc. R News 6:7-11
Podlich DW, Cooper M (1998) QU-GENE: A simulation platform for quantitative analysis of genetic models. Bioinformatics 14:632-653
Prigge V, Maurer HP, Mackill DJ, Melchinger AE, Frisch M (2008) Comparison of the observed with the simulated distributions of the parental genome contribution in two marker-assisted backcross programs in rice. Theor Appl Genet 116:739-744
R Development Core Team (2011) R: A language and environment for statistical computing, Vienna, Austria
Rasmussen CE (2000) The infinite Gaussian mixture model. In advances in neural information processing systems 12. MIT Press, pp 554-560
Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci USA 98:11479-11484
Ribaut J-M, Jiang C, Hoisington D (2002) Simulation experiments on efficiencies of gene introgression by backcrossing. Crop Sci 42:557-565
Robert CP, Casella G (2004) Monte Carlo Statistical Methods (second edition). Springer-Verlag, New York
Santiago E, Caballero A (1995) Effective size of populations under selection. Genetics 139:1013-1030
Schon C, Utz H, Groh S, Truberg B, Openshaw S, Melchinger A (2004) Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. Genetics 167:485
Schulz-Streeck T, Ogutu J, Piepho H-P (2011) Pre-selection of markers for genomic selection. BMC Proc 5:S12
Senior ML, Chin ECL, Lee M, Smith JSC, Stuber CW (1996) Simple sequence repeat markers developed from maize sequences found in the Genbank database: Map construction. Crop Sci 36:1676-1683

Song XF, Song TM, Dai JR, Rocheford TR, Li JS (2004) QTL mapping of kernel oil concentration with high-oil maize by SSR markers. Maydica 49:41-48
Stuber CW, Edwards MD, Wendel JF (1987) Molecular marker-facilitated investigations of quantitative trait loci in maize. II. Factors influencing yield and its component traits. Crop Sci 27:639-648
Sturtevant AH (1915) The behavior of the chromosomes as studied through linkage. Mol Gen Genet 13:234-287
Sun X, Marza F, Ma H, Carver B, Bai G (2010) Mapping quantitative trait loci for quality factors in an inter-class cross of US and Chinese wheat. Theor Appl Genet 120:1041-1051
Sun X, Peng T, Mumm R (2011) The role and basics of computer simulation in support of critical decisions in plant breeding. Molecular Breeding 28:421-436
Timmermans M, Das OP, Messing J (1996) Characterization of a meiotic crossover in maize identified by a restriction fragment length polymorphism-based method. Genetics 143:1771-1783
Tinker NA, Mather DE (1993) GREGOR: Software for genetic simulation. J Hered 84:237
VanRaden P, Van Tassell C, Wiggans G, Sonstegard T, Schnabel R, Taylor J, Schenkel F (2009) Reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92:16 - 24
Visscher PM, Haley CS, Thompson R (1996) Marker-assisted introgression in backcross breeding programs. Genetics 144:1923-1932
Von Korff M, Léon J, Pillen K (2010) Detection of epistatic interactions between exotic alleles introgressed from wild barley (*h. Vulgare* ssp. *Spontaneum*). Theor Appl Genet 121:1455-1464
Wahba G (1990) Spline Models for Observational Data. SIAM, Philadelphia
Wang HW, Han J, Sun WT, Chen SJ (2010) Genetic analysis and QTL mapping of stalk digestibility and kernel composition in a high-oil maize mutant (*zea mays* l.). Plant Breeding 129:318-326
Wang J, Chapman SC, Bonnett DG, Rebetzke GJ, Crouch J (2007) Application of population genetic theory and simulation models to efficiently pyramid multiple genes via marker-assisted selection. Crop Sci 47:582-590
Wang J, Eagles HA, Trethowan R, Van Ginkel M (2005) Using computer simulation of the selection process and known gene information to assist in parental selection in wheat quality breeding. Aus J Agric Res 56:465-473
Wang J, van Ginkel M, Podlich D, Ye G, Trethowan R, Pfeiffer W, DeLacy IH, Cooper M, Rajaram S (2003) Comparison of two breeding strategies by computer simulation. Crop Sci 43:1764-1773
Wang J, van Ginkel M, Trethowan R, Ye G, DeLacy I, Podlich D, Cooper M (2004) Simulating the effects of dominance and epistasis on selection response in the CIMMYT wheat breeding program using QuCim. Crop Sci 44:2006-2018
Wang L, Wang A, Huang X, Zhao Q, Dong G, Qian Q, Sang T, Han B (2011) Mapping 49 quantitative trait loci at high resolution through sequencing-based genotyping of rice recombinant inbred lines. Theor Appl Genet 122:327-340
Watson G (1964) Smooth regression analysis. Sankhya A 26:359-372
Wong CK, Bernardo R (2008) Genomewide selection in oil palm: Increasing selection gain per unit time and cost with small populations. Theor Appl Genet 116:815-824

Wright S (1965) The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution 19:395-420
Wright S (1978) Evolution and the Genetics of Populations, Vol 4: Variability within and among natural populations. University of Chicago Press, Chicago, Illinois
Xiao YN, Li XH, George ML, Li MS, Zhang SH, Zheng YL (2005) Quantitative trait locus analysis of drought tolerance and yield in maize in China. Plant Mol Biol Reporter 23:155-165
Xu S (2003a) Estimating polygenic effects using markers of the entire genome. Genetics 163:789-801
Xu S (2003b) Theoretical basis of the Beavis effect. Genetics 165:2259-2268
Xu S, Jia Z (2007) Genomewide analysis of epistatic effects for quantitative traits in barley. Genetics 175:1955-1963
Xu Y (2010) Molecular Plant Breeding. CABI, Cambridge, MA
Xu Y, Crouch JH (2008) Marker-assisted selection in plant breeding: From publications to practice. Crop Sci 48:391-407
Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539-551
Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203-208
Zeng Z-B, Wang T, Zou W (2005) Modeling quantitative trait loci and interpretation of models. Genetics 169:1711-1725
Zeng ZB (1992) Correcting the bias of Wright's estimates of the number of genes affecting a quantitative character: A further improved method. Genetics 131:987-1001
Zhao H, Speed T, McPeek M (1995) Statistical analysis of crossover interference using the chi-square model. Genetics 139:1045-1056
Zhong S, Jannink JL (2007) Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. Genetics 177:567-576